Создание цифровой библиотеки коллекций периодических изданий на основе Greenstone

В.А. Резниченко, Г.Ю. Проскудина, О.М. Овдей
Институт программных систем НАН Украины

Аннотация

В работе дано описание процесса создания цифровой библиотеки научных периодических изданий на основе программного обеспечения Greenstone. Кроме того, представлена информационная мо-дель научного периодического издания; определены понятия цифровая библиотека и коллекция; приведен краткий обзор программного обеспечения Greenstone, а также возможностей этой системы по созданию коллекций цифровой библиотеки на примере научного журнала.

1 Введение

Создание коллекций информационных ресурсов – одно из важных направлений разработки цифровых библиотек (ЦБ) [1]. Коллекции представляют собой наиболее распространенную форму организации инфор-мационных ресурсов в ЦБ [2]. В связи с широкими возможностями существующих информационных техно-логий и разнообразием природы информационных ресурсов, характеристики коллекций весьма многообраз-ны. Однако коллекции обладают некоторыми общими свойствами, понимание которых имеет существенное значение при их разработке.

В данной работе для создания ЦБ научных периодических изданий предлагается использовать программ-ное обеспечение Greenstone - эффективное Open Source-решение для построения ЦБ [3-5]. Система обеспе-чивает поиск с предварительным индексированием по документам всех популярных форматов и, прежде всего doc и pdf, которые могут быть представлены в заархивированном виде. Система создает каталог доку-ментов, конвертирует их в xml-формат, а затем обеспечивает отдаленный доступ к библиотеке посредством браузера.

Предлагаемая работа содержит: описание информационной модели научного периодического издания (часть 2); краткий обзор программного обеспечения (ПО) Greenstone (часть 3); обсуждение его возможно-сти по созданию коллекций ЦБ на примере научного журнала (часть 4), а также перспектив развития систе-мы Greenstone (часть 5).

2 Описание предметной области

Предметом нашего анализа в данной статье является научное периодическое издание. В общем случае такое издание можно представить информационными ресурсами следующих трех уровней: периодическое издание в целом, выпуск периодического издания и отдельная публикация в выпуске.

2.1 Периодическое издание

Периодическое издание имеет следующие основные характеристики или атрибуты:

1. *Название*. Имя, данное периодическому изданию. Может иметь несколько значений на разных язы-ках.
2. *Тематические разделы и подразделы*. Полный перечень тем содержания, выражающийся с помо-щью ключевых фраз, ключевых слов или классификационных кодов, которые описывают темы пуб-ликаций. На практике значения выбираются из контролируемого словаря или формальной схемы классификации, например, УДК. Тематические разделы могут иметь структуру простого линейного списка или иерархически структурированными.
3. *Описание*. Краткое сообщение о содержании периодического издания. Описание может быть пред-ставлено в виде: реферата, содержания, ссылки на графическое представление или простого тексто-вого изложения содержания.
4. *Тип*. Характер или жанр содержания периодического издания.
5. *Издатель*. Организация, ответственная за периодическое издание.
6. *Главный редактор*.
7. *Редакционная коллегия*.
8. *Дата*. Содержит дату основания периодического издания.
9. *Периодичность*.
10. *Идентификаторы* ISBN (International Standard Book Number), ISSN (International Standard Serial Number), DOI (Digital Object Identifier). Возможно несколько значений.
11. *Язык*. Содержит язык интеллектуального содержания периодического издания.
12. *Адрес*. Содержит адрес редакции периодического издания.
13. *Телефон*.
14. *E-mail*.
15. *URL*. Может иметь несколько значений.

2.2 Выпуск

Описывает конкретный выпуск периодического издания и имеет следующие атрибуты:

1. *Тема выпуска*. Возможно несколько значений.
2. *Посвящение*. Например, "К 80-летию со дня рождения В.М. Глушкова".
3. *Дата и номер*.
4. *Содержание*. Может иметь несколько значений на разных языках.

2.3 Публикация

Описывает конкретную публикацию (статью) в выпуске журнала. Имеет следующие атрибуты:

1. *Название*. Несколько значений на разных языках.
2. *Автор*. Может иметь несколько значений.
3. *Тип публикации*. Например, статья, сообщение, доклад на конференции.
4. *Библиографическое описание*.
5. *Язык*.
6. *Полный код УДК*. (Международная система тематической классификации публикаций, UDC - Universal Decimal Classification) [6].
7. *Специальные идентификаторы*. Например, идентификаторы этой публикации в системе ACM Classification System [7].
8. *Дата* принятия статьи.
9. *Реферат*. Несколько значений, возможно, на разных языках.
10. *Ключевые слова*, характеризующие содержание публикации. Несколько значений на разных языках.
11. *Номера страниц*.
12. *Полный текст*. Может быть представлен файлом или URL.

Пример описания публикации.

Название: “Онтологии в контексте интеграции информации: представления, методы и инструменты построения”.

Автор: Овдий О.М., Проскудина Г.Ю.

Тип публикации: Статья.

Библиографическое описание: Овдий О.М., Проскудина Г.Ю. Онтологии в контексте интеграции информации: представления, методы и инструменты построения // Проблемы программирования. ? 2004. ? №2-3 ? С.353-365.

Язык: украинский.

Полный код УДК: 004.82

Название периодического издания: Проблемы программирования.

Дата и номер: 2004, №2-3.

Реферат: Рассматривается использование онтологий для поддержки задач интеграции в семантически гетерогенных информационных системах. Представлены основные понятия и определения онтологий, цели и примеры их построения.

Ключевые слова, характеризующие содержание публикации: онтология, построение онтологий, объединение онтологий, инженерия онтологий, отображение онтологий.

Далее рассмотрим, как можно создать информационную систему, которая относится к классу *цифровых библиотек*, построив коллекцию информационных ресурсов научное издание - выпуск – публикация согласно приведенной выше модели. Коллекцию будем строить на основе ПО Greenstone.

3 Краткое описание Greenstone

Greenstone - комплексная система для построения и распространения коллекций ЦБ. Она обеспечивает способ организации и публикации информации в Интернете (или на CD-дисках). Следовательно, система Greenstone может решить задачу сохранения и извлечения в электронном виде периодических изданий и удовлетворить потребность научных работников в получении информации о периодическом издании, выпуске периодического издания или публикации.

ПО Greenstone разработано на факультете компьютерных наук университета Вайкато в Новой Зеландии в рамках проекта по созданию цифровых библиотек. Руководитель проекта – Ян Виттен (Ian H. Witten). Разработка проводилась при содействии ЮНЕСКО и неправительственной организации Human info. Распространяется с ноября 2000 года. В настоящее время Greenstone постоянно дорабатывается. Программа свободно доступна на сайте http://greenstone.org и отвечает условиям GNU.

Существует две версии Greenstone – локальная и сетевая. Система работает на платформах Windows и Unix с использованием стандартных Web-серверов.

В настоящее время Greenstone широко используется многими организациями разных стран. На упомянутом выше сайте имеются ссылки на более чем 20 коллекций цифровых библиотек Greenstone. На сайте http://www.nzdl.org можно посмотреть более 50 коллекций ЦБ, созданных при содействии разработчиков системы. Показательные коллекции включают статьи из газет, технические документы, художественные книги, научные журналы, фольклор, аудио и видео информацию.

3.1 Функции и возможности Greenstone

ПО Greenstone предоставляет возможности [8]:

- создавать *коллекции* электронных документов;
- детально определять документы в зависимости от метаданных;
- сохранять десятки Гб текста и связанных с ним изображений;
- осуществлять полнотекстовый поиск, а также поиск и просмотр документов по полям метаданных;
- документы, которые вносятся в коллекцию, и их метаданные могут иметь разные форматы;
- осуществлять обработку документов на каком-либо языке и поддерживать многоязычный интерфейс пользователя;
- организовывать и публиковать информацию в Интернете или на компакт-дисках;
- использовать стандартные и нестандартные метаданные для описания содержания документов.

Далее, рассматривая Greenstone, остановимся на некоторых ключевых моментах.

3.2 Коллекции

ЦБ, созданная с помощью Greenstone, содержит множество коллекций, организованные по отдельности они имеют много сходства. Легко поддерживаемые, эти коллекции могут быть дополнены и автоматически перестроены.

Коллекции - совокупность документов разных форматов, собранных вместе на основе обусловленных пользователем критериев и к которым применяются единые механизмы сохранения, индексации, поиска, просмотра и представления.

Коллекции могут состоять из сотен тысяч и даже миллионов документов. Коллекции могут включать документы разной природы: текстовые документы (статьи, журналы, газеты, отчеты), а также аудио и видео-документы. В коллекции можно создавать подколлекции, и в некоторых случаях, коллекции можно логически объединять.

Каждый текстовый документ может быть иерархически структурирован в виде вложенных разделов (sections) (разделы, подразделы, подподразделы и т.д.). Иерархическая структура разделов отображает содержательную структуру документа. Каждый из разделов, в свою очередь, состоит из одного или нескольких абзацев (paragraphs). Таким образом, структуризация содержания обычных документов на части, главы, разделы и т.д. представляется в документах Greenstone в виде иерархической структуры разделов Greenstone. Структура документа может использоваться при формировании поисковых индексов. Если входные документы не имеют структуры, то в коллекции Greenstone они могут быть представлены в виде последовательности страниц, что позволяет просматривать документы постранично.

Входные информационные ресурсы для построения коллекции могут располагаться: на локальном компьютере, в локальной сети и глобальной сети и доступны с использованием протоколов HTTP и FTP.

Входные документы могут иметь разные форматы, для поддержки импорта которых используются плагины (специальные утилиты импорта документов соответствующих форматов). Все входные документы, внесенные в систему Greenstone, конвертируются в формат архива Greenstone (Greenstone Archive Format). Система Greenstone каждому документу автоматически присваивает уникальный идентификатор OID (Object Identifier).

В Greenstone структура каждой коллекции определяется в процессе ее создания. Она включает определение формата используемых документов, их вывод на экран, источник метаданных, какие предметные показатели должны быть включены, какие следует предоставить полнотекстовые индексы, как должны отображаться результаты поиска. После того, как коллекция создана, в нее легко добавить новые документы при условии, что они того же формата, что и существующие документы, и что они имеют сходные метаданные. Каждая коллекция содержит файл конфигурации [5], в котором устанавливаются параметры построения и использования коллекции.

На Рис.1 представлена домашняя страница коллекции журнала, которая содержит общую информацию о коллекции (см. п. 2.1).

Коллекции можно открыть для *поиска* и *просмотра*.

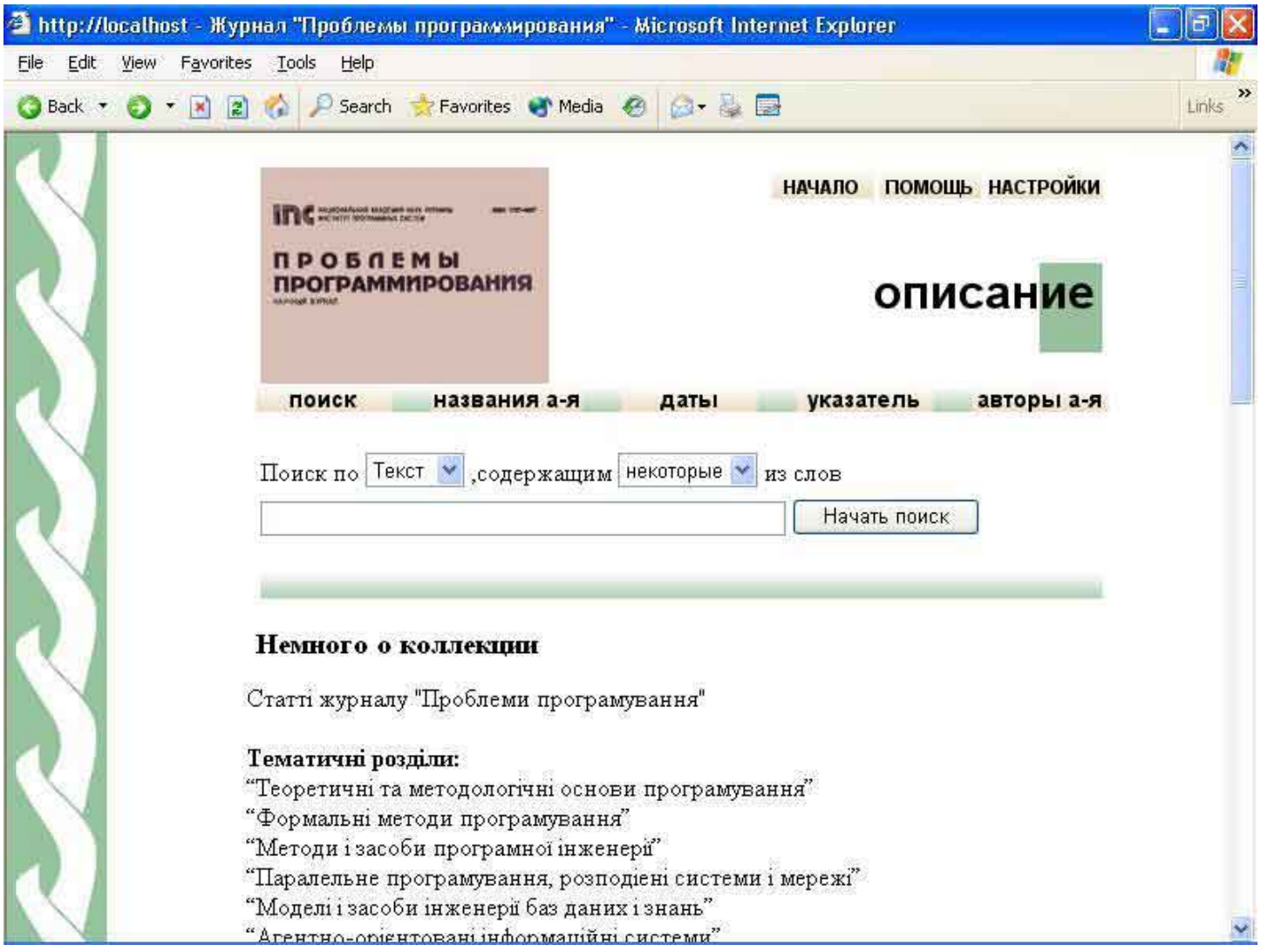

Рис. 1 Домашняя страница коллекции

3.3 Поиск

Пользователь Greenstone может осуществлять полнотекстовый поиск. Диапазон поиска определяют индексы, которые строятся на разных частях документов. С помощью индексов можно искать по отдельному слову, набору слов или фраз. Коллекции могут иметь индексы полных документов, индексы параграфов, индексы определенных метаданных (например, названий или авторов) по каждому из которых можно осуществлять поиск определенных слов или фраз. Результаты могут быть упорядочены или отсортированы по элементам метаданных. Greenstone предоставляет возможность выполнять поиск по нескольким коллекциям сразу с последующим объединением результатов поиска.

На Рис.2-3 показаны экраны выполнения поиска в коллекции соответственно – формулировка запроса и результат.

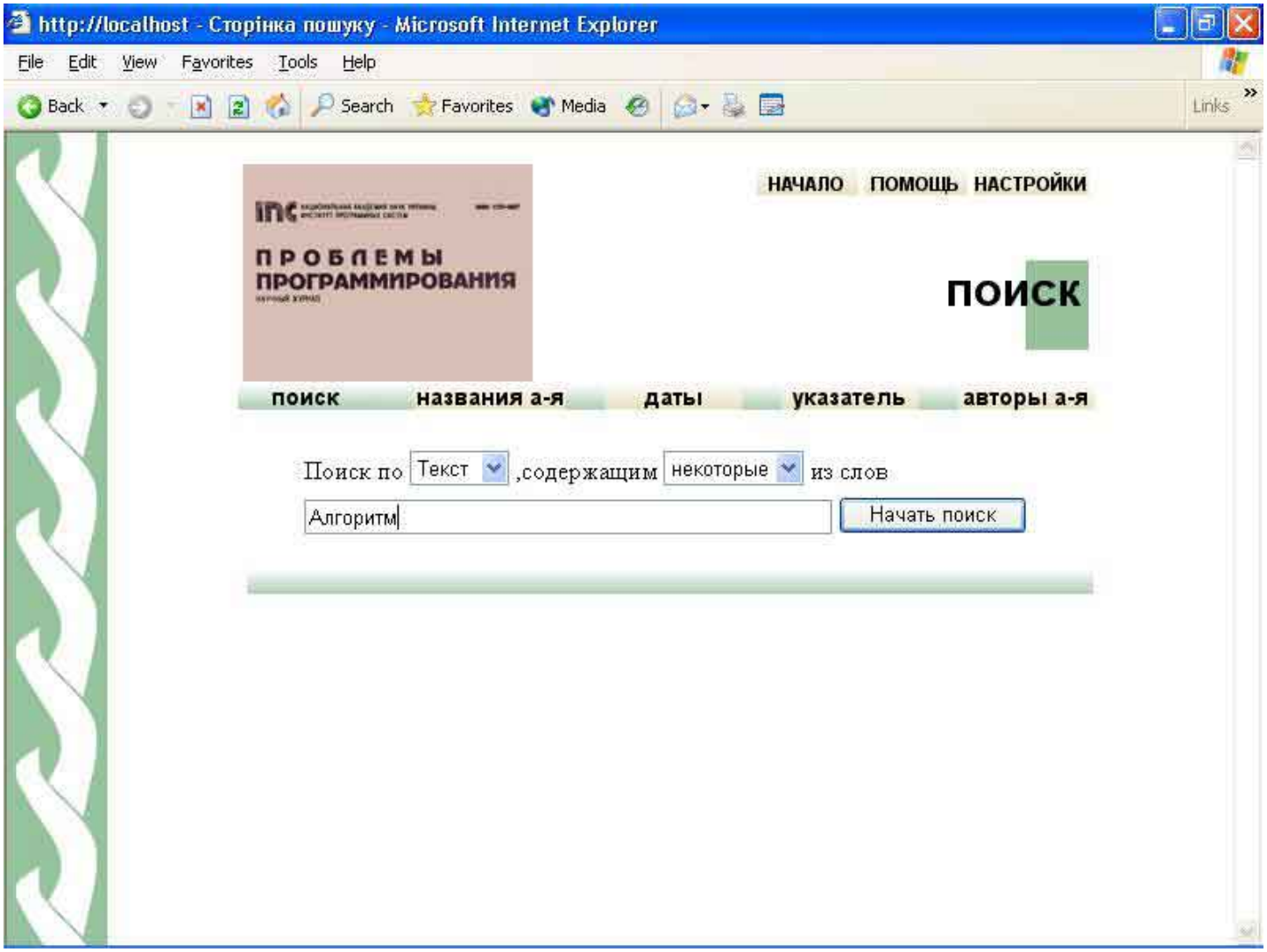

Рис. 2 Запрос на поиск в коллекции

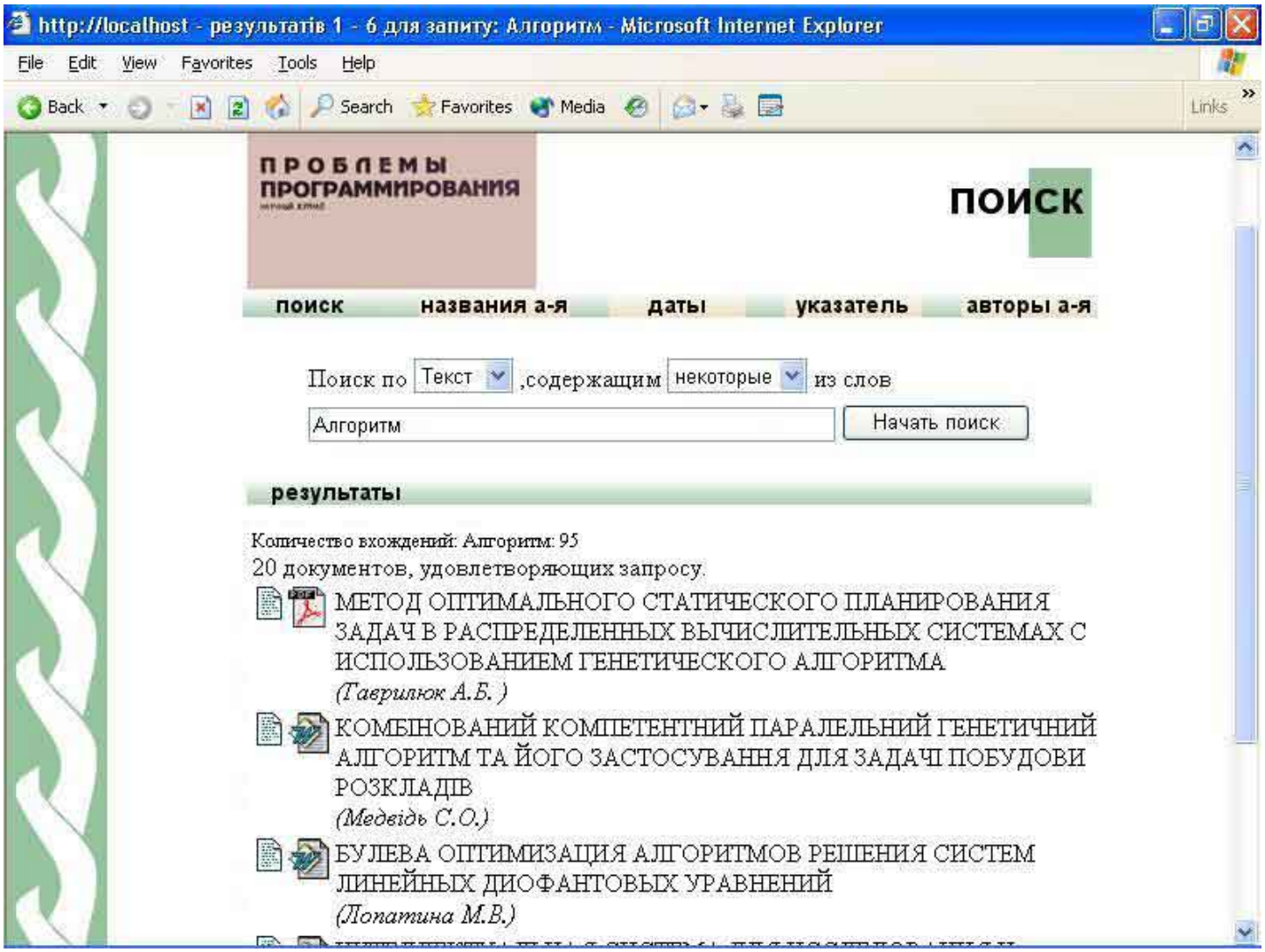

Рис. 3 Результат выполнения команды поиска

3.4 Просмотр

Для просмотра коллекции используется определенный перечень метаданных: перечень авторов, названий, дат, иерархичные классификационные структуры и т.д. Метаданные являются основой и начальным пунктом для осуществления просмотра. Разные коллекции предлагают разные возможности для просмотра. Интерфейсы просмотра и поиска создаются в процессе построения коллекции согласно информации о конфигурации коллекции.

Для создания структур просмотра метаданных, используется система классификаторов. С их помощью можно создать индексы просмотра такие как: алфавитные показатели, данные и разнообразные иерархические структуры. Можно создавать новые структуры просмотра.

В Greenstone разработан набор стандартных классификаторов [5]. Все классификаторы генерируют иерархическую структуру, используемую для отображения индекса просмотра. На самом нижнем уровне этой структуры естественно размещаются документы, но могут определяться и разделы документов. Классификаторы могут иметь установленное или произвольное число уровней иерархии.

На Рис.4-10 демонстрируется процесс просмотра коллекции журнала.

Рис. 4 Просмотр коллекции по названиям статей

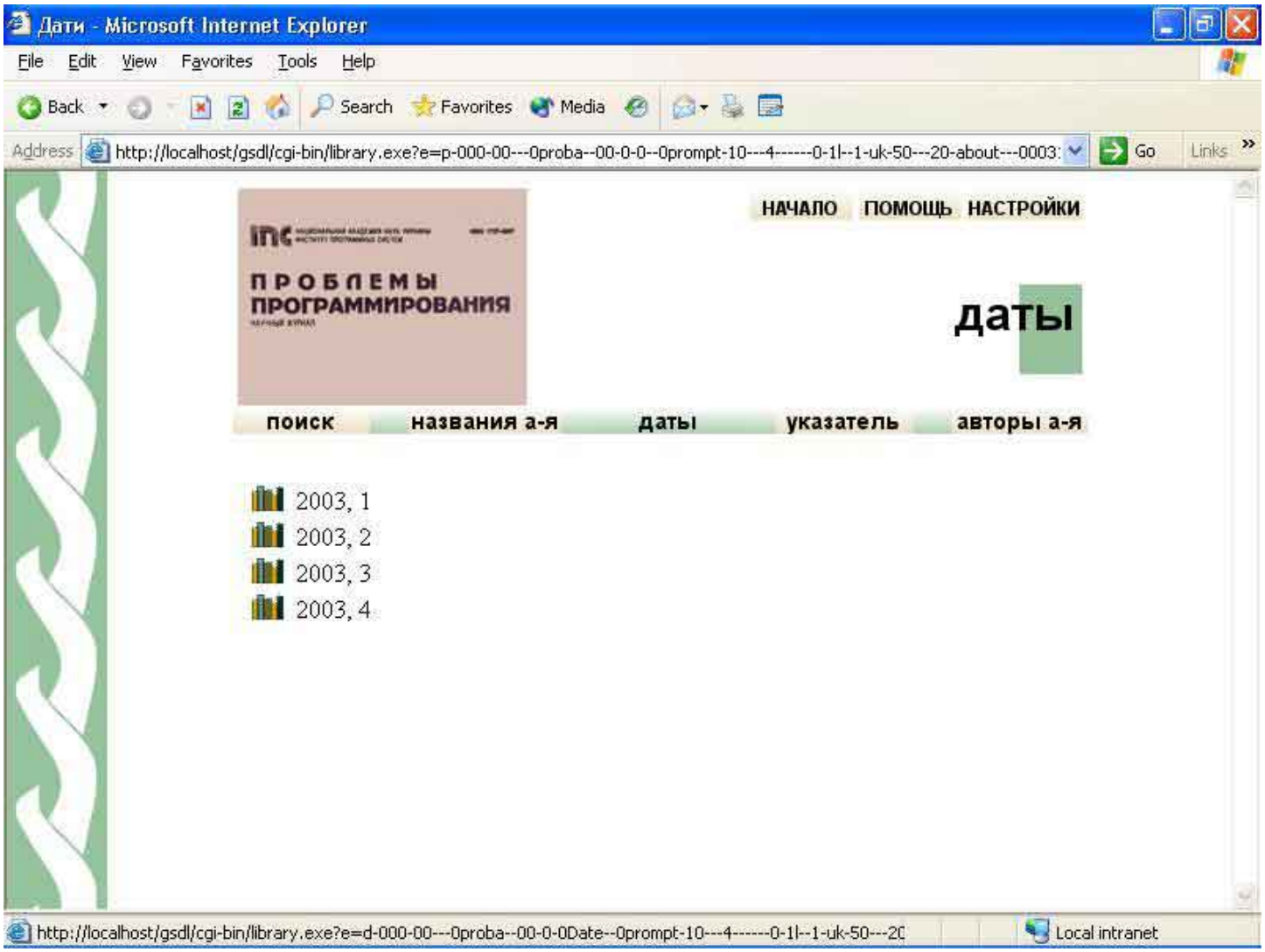

Рис. 5 Просмотр коллекции по выпускам

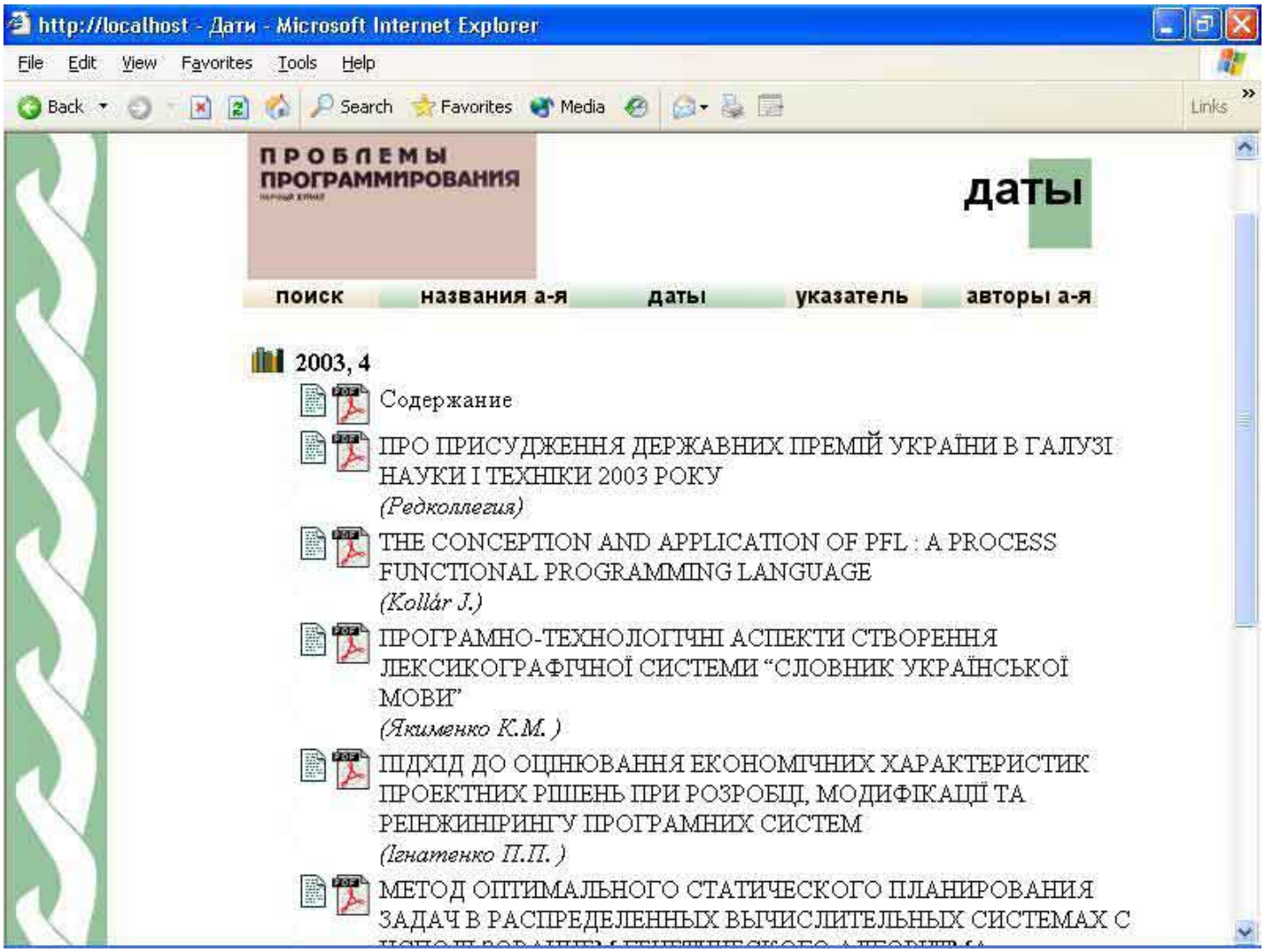

Рис. 6 Просмотр одного выпуска.

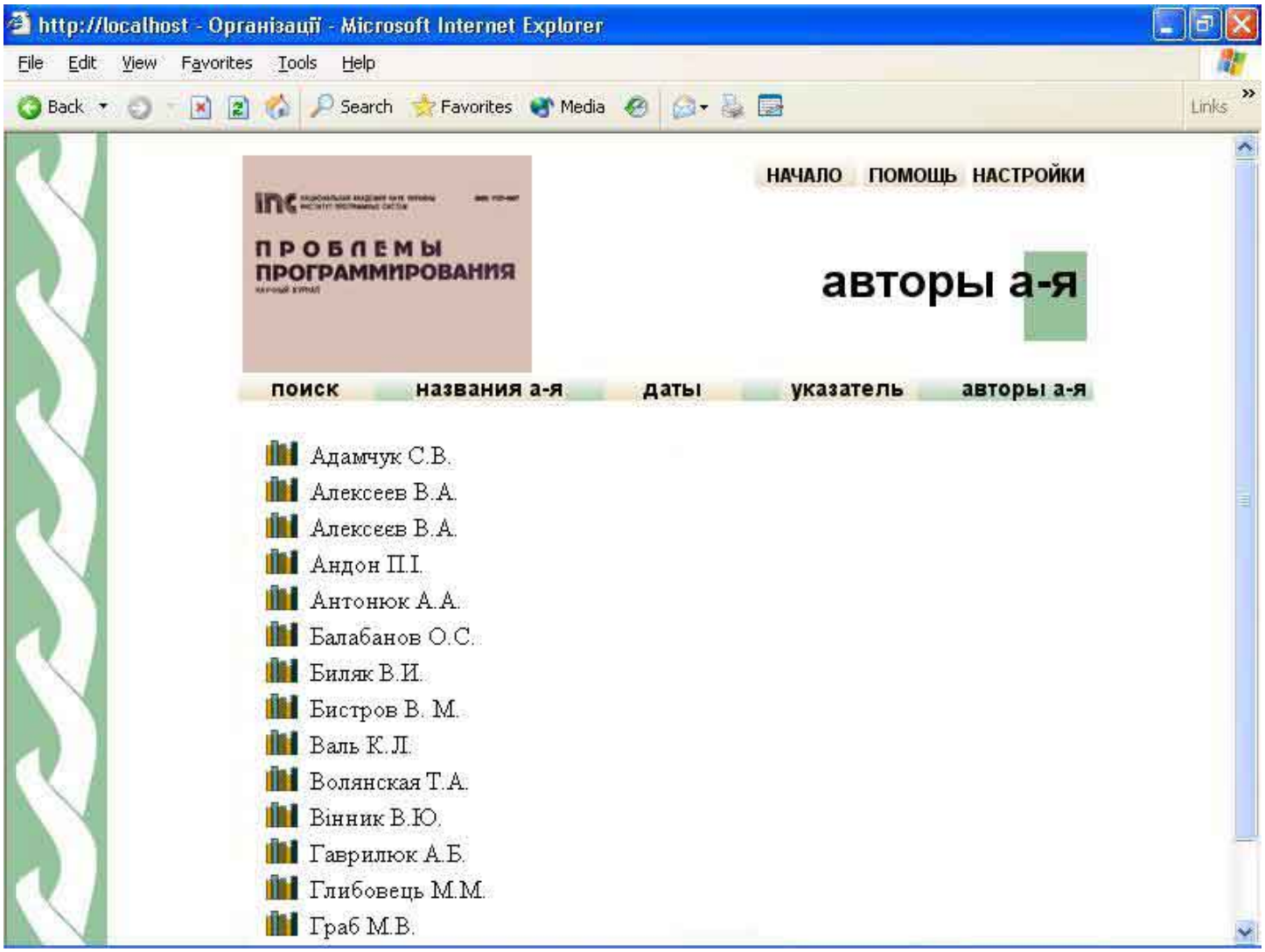

Рис. 7 Просмотр коллекции по авторам публикаций

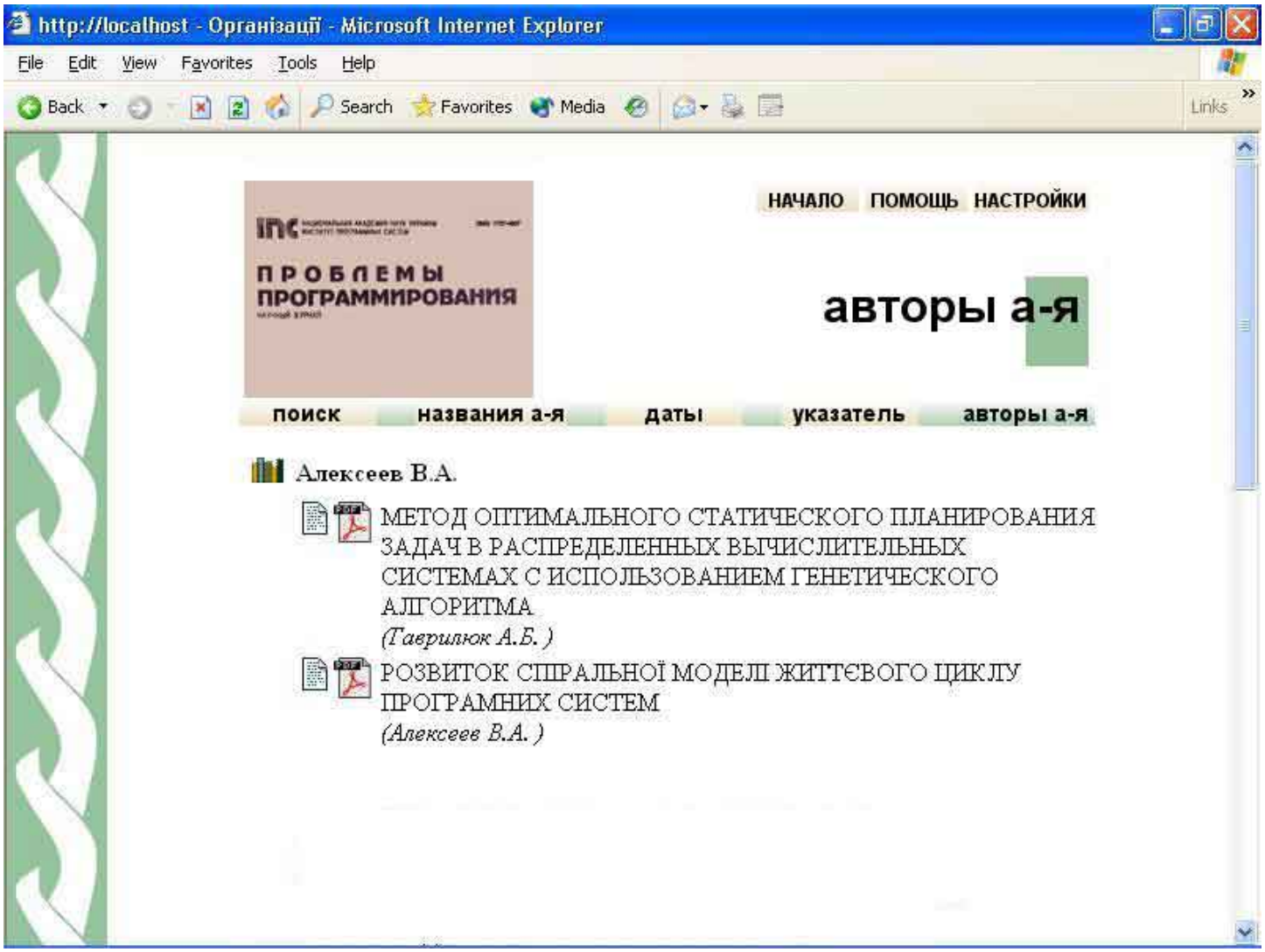

Рис. 8 Просмотр публикаций одного автора

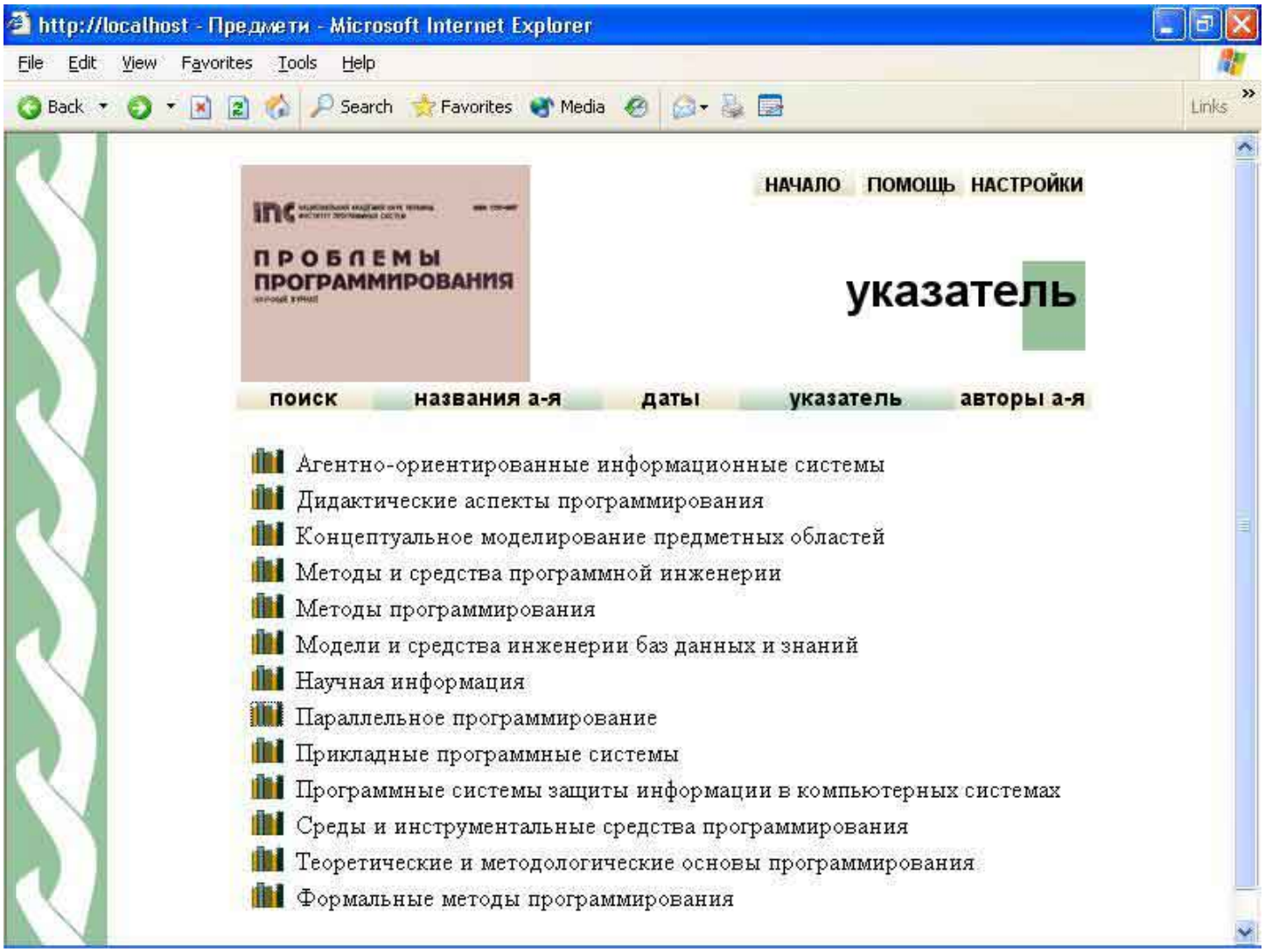

Рис. 9 Просмотр по предметной классификации

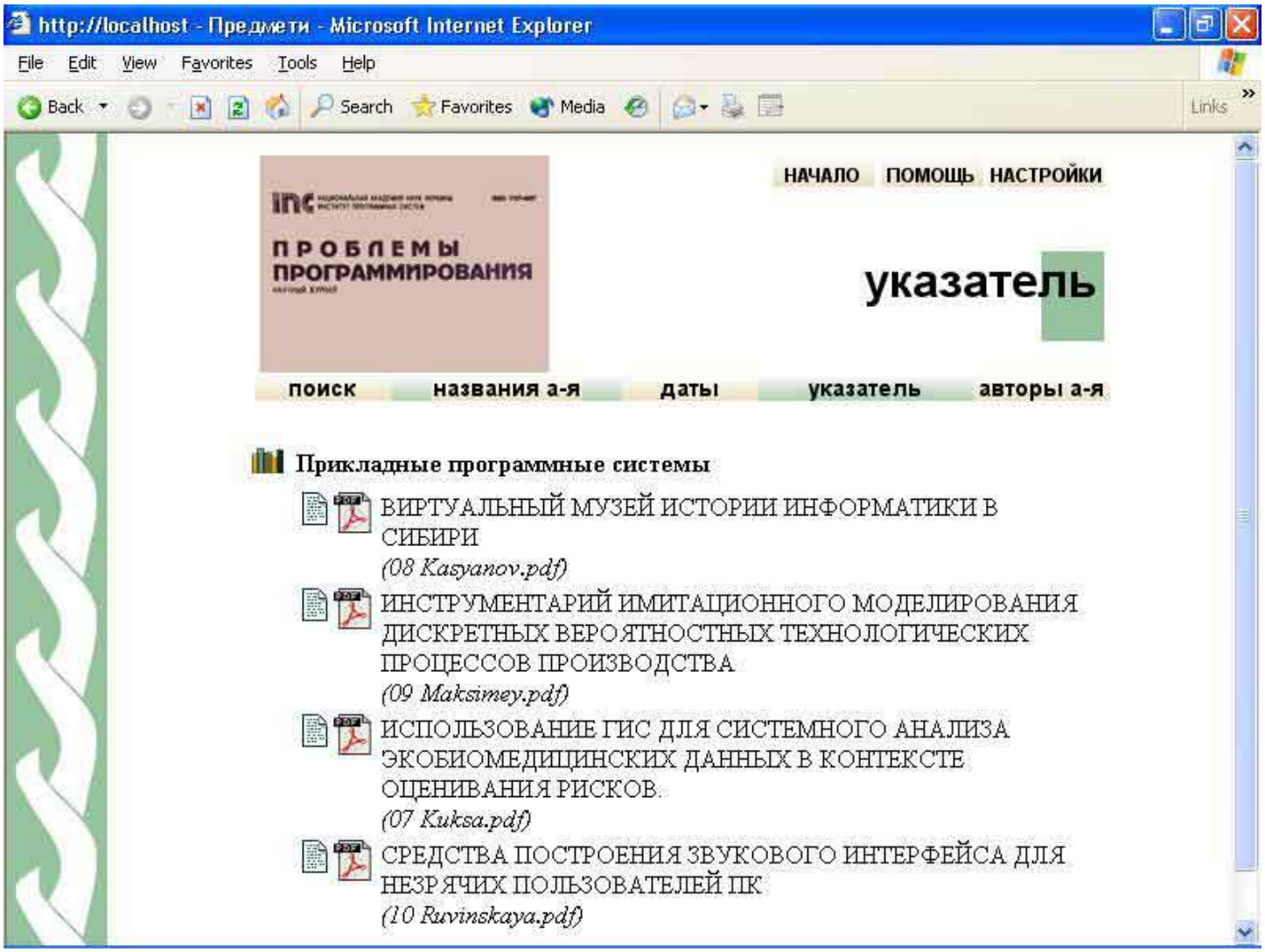

Рис. 10 Просмотр по предметной классификации

3.5 Многоязычность

В системе используется набор символов UNICODE. В связи с этим и документы, и внешний интерфейс могут представляться на разных языках. В этом смысле система Greenstone является многоязычной. Кроме того, систему легко расширить новым языком интерфейса, путем добавления соответствующих названий и описаний элементов интерфейса на желаемом языке в файлы конфигурации.

3.6 Дополнение

Следует также добавить, что возможности, обеспечиваемые Greenstone, и интерфейс, откуда пользователи цифровой библиотеки обращаются к ним, легко настраивать на разных уровнях. Можно задавать форматы документов (например, HTML, Word, PDF, Postscript, PowerPoint, Exel) или формат изображений (например, TIFF, GIF, PNG, JPEG), включаемых в коллекцию. Кроме того, можно задавать набор доступных метаданных (например, MARC, архивы OAI, BibTex, базы данных CDS/ISIS), а также какие будут обеспечиваться полнотекстовые индексы (например, всего текста, возможно, разделенные по языку или другими признаками, и выбранными метаданными, например, заголовками или резюме) и структуры просмотра (например, список авторов, заголовков, дат, иерархия предметной классификации). Все это осуществляется с помощью интерфейса библиотекаря GLI (Greenstone's Librarian Interface).

4 Создание коллекций научных периодических изданий с помощью GLI

Создание коллекций проводится с помощью интерфейса библиотекаря GLI, инструмента сбора и обработки документов с последующим созданием коллекций цифровых библиотек, работающих под управлением Greenstone. Он обеспечивает доступ к функциональным возможностям ПО библиотеки Greenstone графическим путем. GLI позволяет добавлять документы и метаданные в коллекцию, создавать новые коллекции и настраивать их на удобный просмотр.

4.1 Роль и структура метаданных

Организация цифровой библиотеки главным образом опирается на метаданные — структурированную информацию о ресурсах (документах), имеющихся в библиотеке. Метаданные это что-то наподобие традиционных карточных каталогов, "кирпич и цемент" библиотек (независимо от того, компьютеризированы они или нет) [9]. Если библиотека "структурирована", то ней можно осмысленно управлять без обязательного понимания ее смысла. Например, для коллекций документов, библиографическая информация о каждом документе была бы метаданными для коллекции. Метаданные документа содержат информацию описательного характера, такую как данные об авторе, заголовок, дату, ключевые слова и т.д. Метаданные могут ассоциироваться с документом в целом или с отдельными разделами документа. Понятие "метаданные" не абсолютное, а относительное: оно только действительно значимо в контексте и ясно дает понять, чем собственно являются данные.

Использование метаданных в качестве строительного материала – действительно определяющая характеристика цифровых библиотек: это

– то, что отличает ее от других коллекций интерактивной информации. Метаданные позволяют расположить в библиотеке новый материал и закрепить за существующими структурами таким образом, что он сразу же становится полноправным членом библиотеки.

Метаданные являются основой для организации индексирования документов, построения классификаторов и также могут использоваться при описании форматов представления результатов поиска или просмотра документов.

В Greenstone с каждой коллекцией связывается один или несколько наборов элементов метаданных. Сейчас существуют десятки таких наборов. Это могут быть узкоспециализированные наборы, предназначенные для описания ресурсов какой-то определенной отрасли или тематики, имеются также и метаданные более универсального характера. Исчерпывающую информацию по наборам метаданных можно получить по адресу: http://www.ifla.org/II/metadata.htm.

GLI в качестве стандартного набора предлагает использовать Дублинское ядро (Dublin Core – DC) [10], который является форматом описания практически любых ресурсов Интернет. Набор DC - несложный по структуре, относительно легкий в применении, расширяемый и интернациональный, т.е. нашедший свое применение по всему миру. В 2001 г. набор элементов метаданных DC был утвержден в США Американским Институтом Национальных Стандартов как стандарт Z39.85 – 2001 (это уже и стандарт ISO 15836-2003).

В ряде стран формат DC рекомендован и принят как государственный стандарт для on-line ресурсов и электронной коммерции [11].

GLI дает возможность определять новые наборы — как правило, добавляя несколько дополнительных элементов к существующему набору. Еще один важный набор - набор извлеченных метаданных, содержащий информацию, автоматически извлеченную непосредственно из документов. Например, для HTML-файлов это содержащиеся метаданные в тэге заголовок, тэге META, или встроенные метаданные в DOC-файлы, автор и заголовок.

Система сохраняет наборы метаданных, используя разные пространства имен (namespaces). Например, документы могут иметь два атрибута Заголовок из набора метаданных Дублинское ядро (dc.Title) и из набора извлеченных метаданных (ex.Title). Они не обязательно должны иметь одинаковые значения. Перечень описательных элементов как для документа в целом, так и его разделов не фиксирован. Документ и его разделы могут содержать свои собственные описательные элементы (т.е. их состав может изменяться от документа к документу или от одного раздела документа к другому). Извлеченные метаданные располагаются непосредственно в документах, а наборы метаданных в отдельных файлах в формате XML. Элементы метаданных имеют вид:

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

Для того чтобы ускорить ручной ввод метаданных, GLI позволяет связывать метаданные как с папками документов, так и с отдельными документами. Это означает, что пользователи могут использовать преимущество группировки документов, чтобы записать общие для группы документов метаданные за одну операцию. Метаданные в Greenstone могут быть простой текстовой строкой (например, название, автор, издатель). Или они могут быть иерархически структурированы. Кроме того, они многозначные, то есть каждый элемент может иметь более одного значения. Это используется, например, когда публикация имеет несколько авторов. GLI сохраняет уже введенные значения метаданных, и там, где нужно их еще раз использовать, они просто выбираются из списков без необходимости их повторного введения.

4.2 Работа с GLI

С помощью GLI пользователи собирают наборы документов, импортируют или описывают метаданные и формируют их в коллекции Greenstone. GLI поддерживает пять основных действий, которые могут чередоваться, но они имеют свой логический порядок.

1. Внесение документов в коллекцию. Документы, импортируемые из существующих коллекций, прибывают с присоединенными метаданными.
2. Обогащение документов путем добавления к ним метаданных.
3. Проектирование коллекции, т.е. определение ее внешнего вида и средств доступа.
4. Построение коллекции с использованием Greenstone.
5. Передача вновь созданной коллекции библиотечному серверу Greenstone. Коллекция автоматически инсталлируется в персональную цифровую библиотеку пользователя, и открывается Web-страница, показывая домашнюю страницу коллекции (Рис.1).

На Рис.11-20 представлена работа GLI на примере построения коллекции журнала «Проблемы программирования». Это экраны в разных точках в течение взаимодействия пользователя и интерфейса, выполняя один проход последовательно по всем указанным выше шагам.

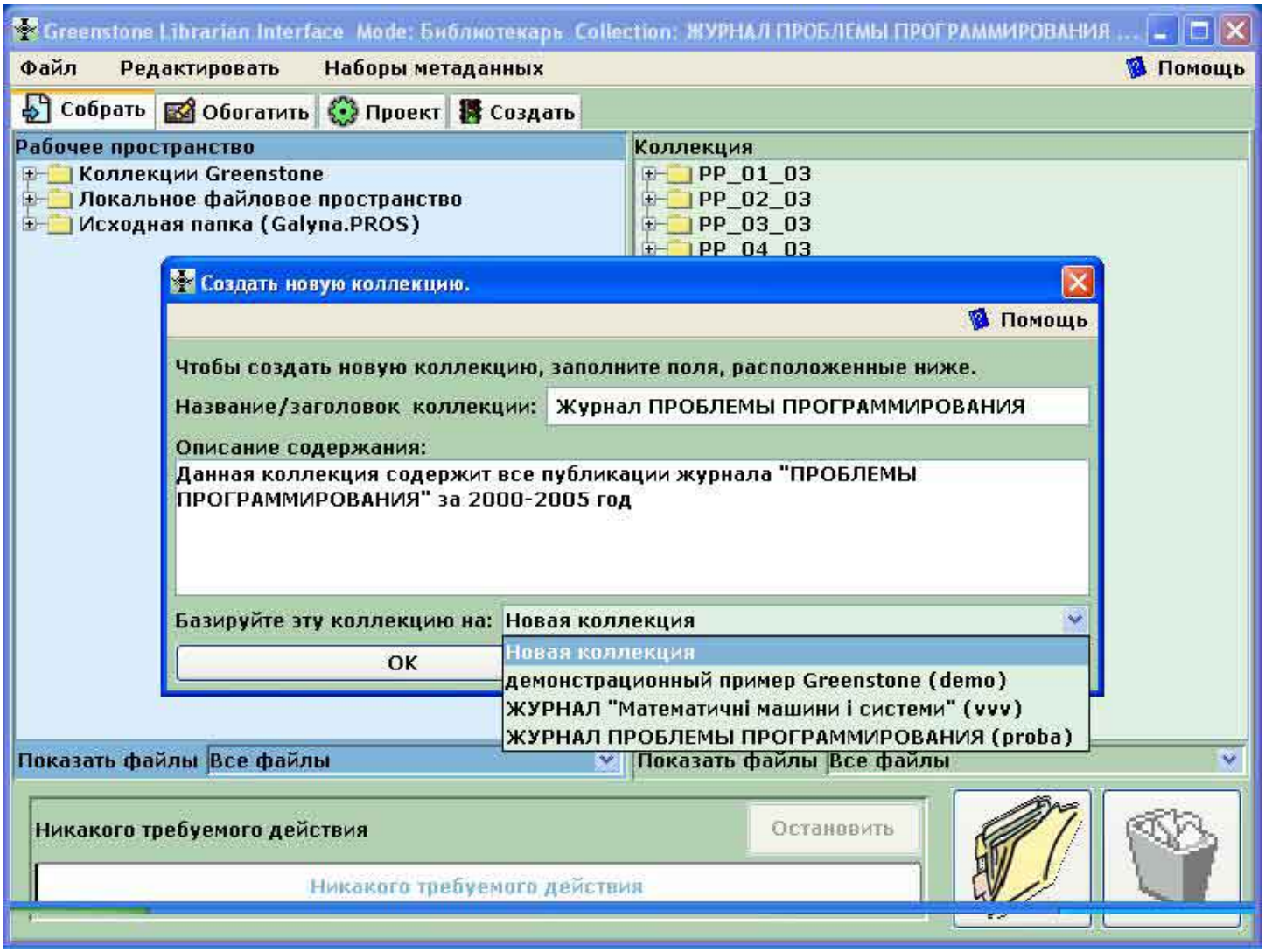

Рис. 11 Запуск новой коллекции

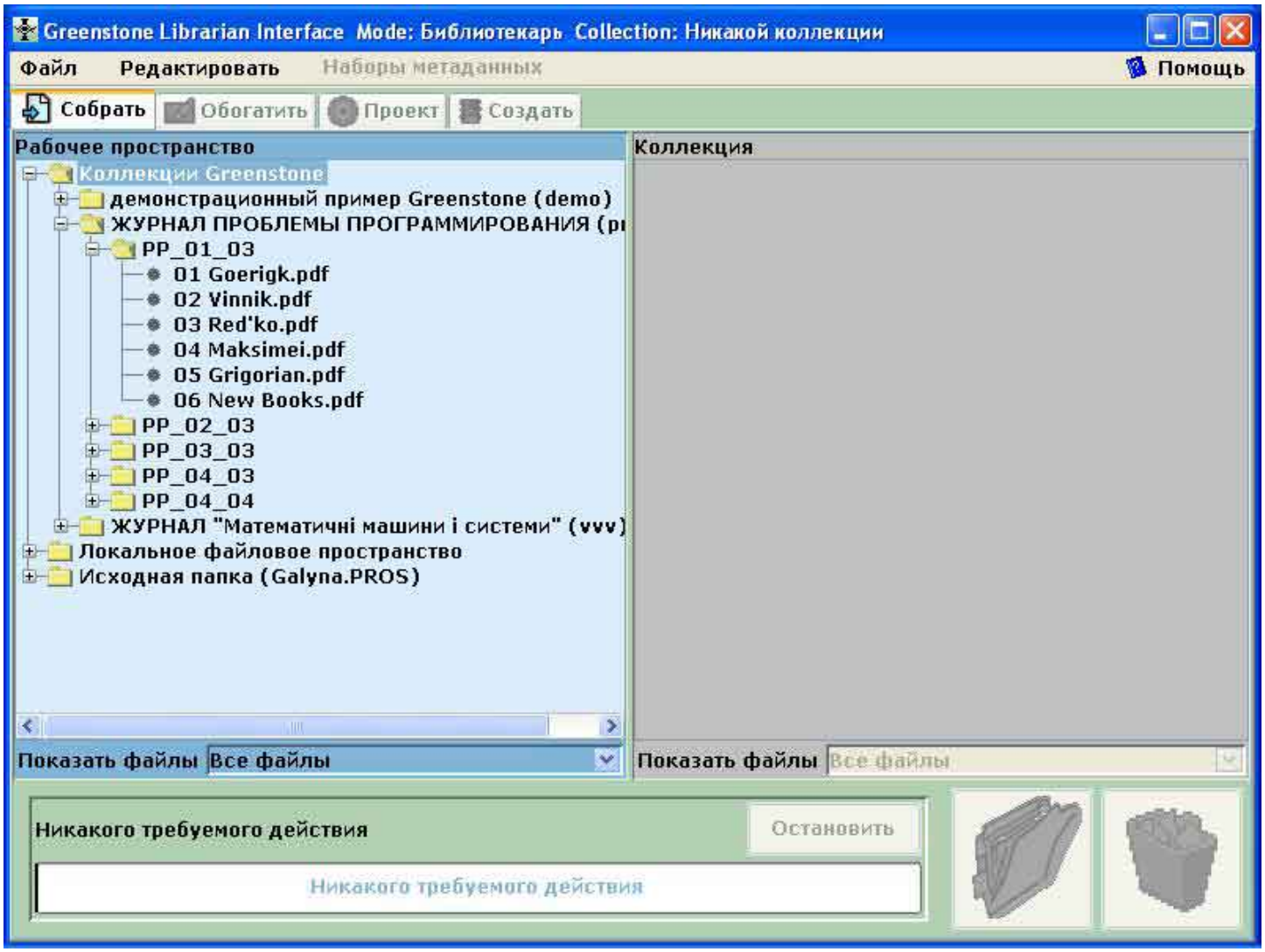

Рис. 12 Исследование локального файлового пространства

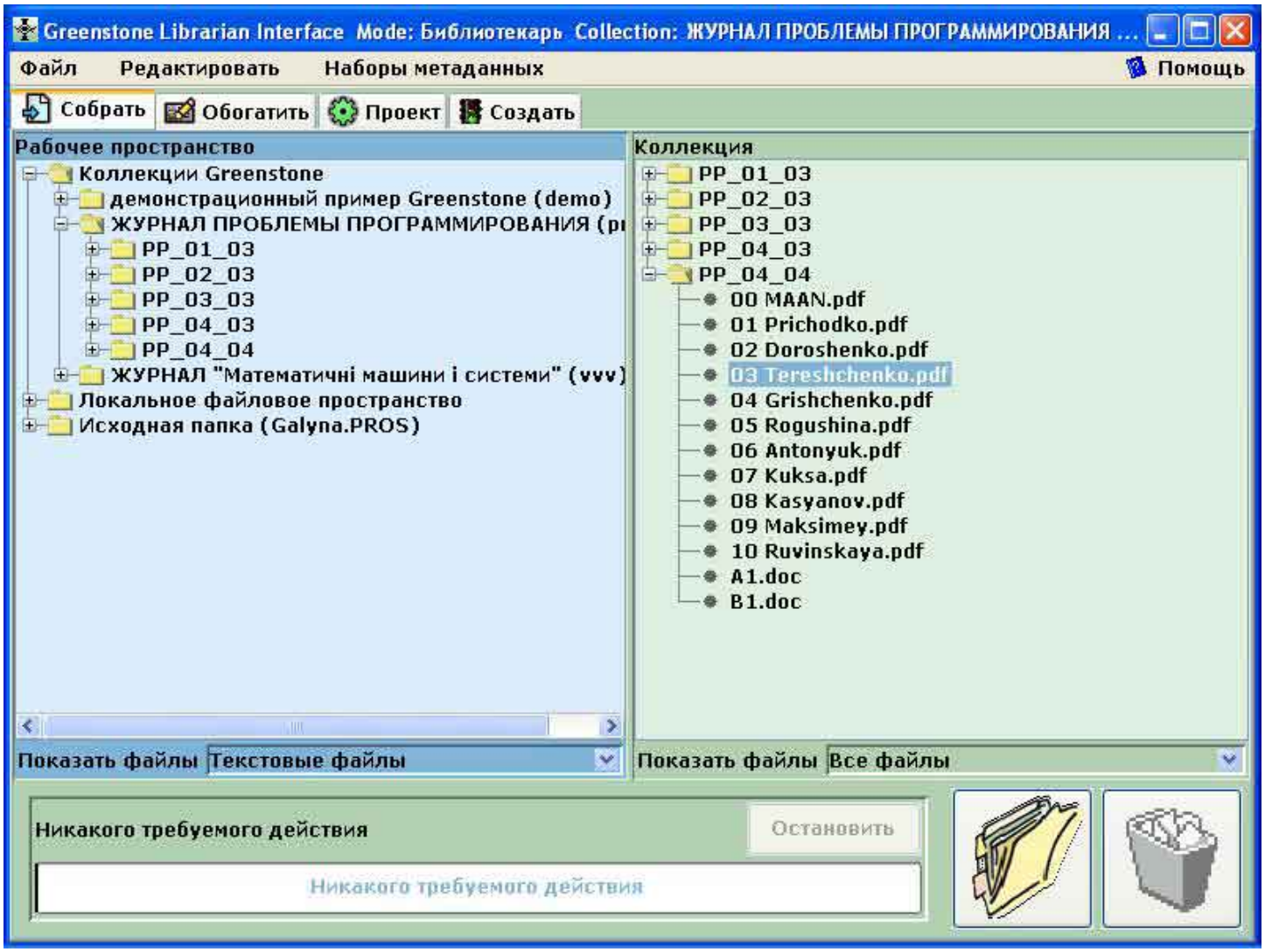

Рис. 13 Фильтрация файлов

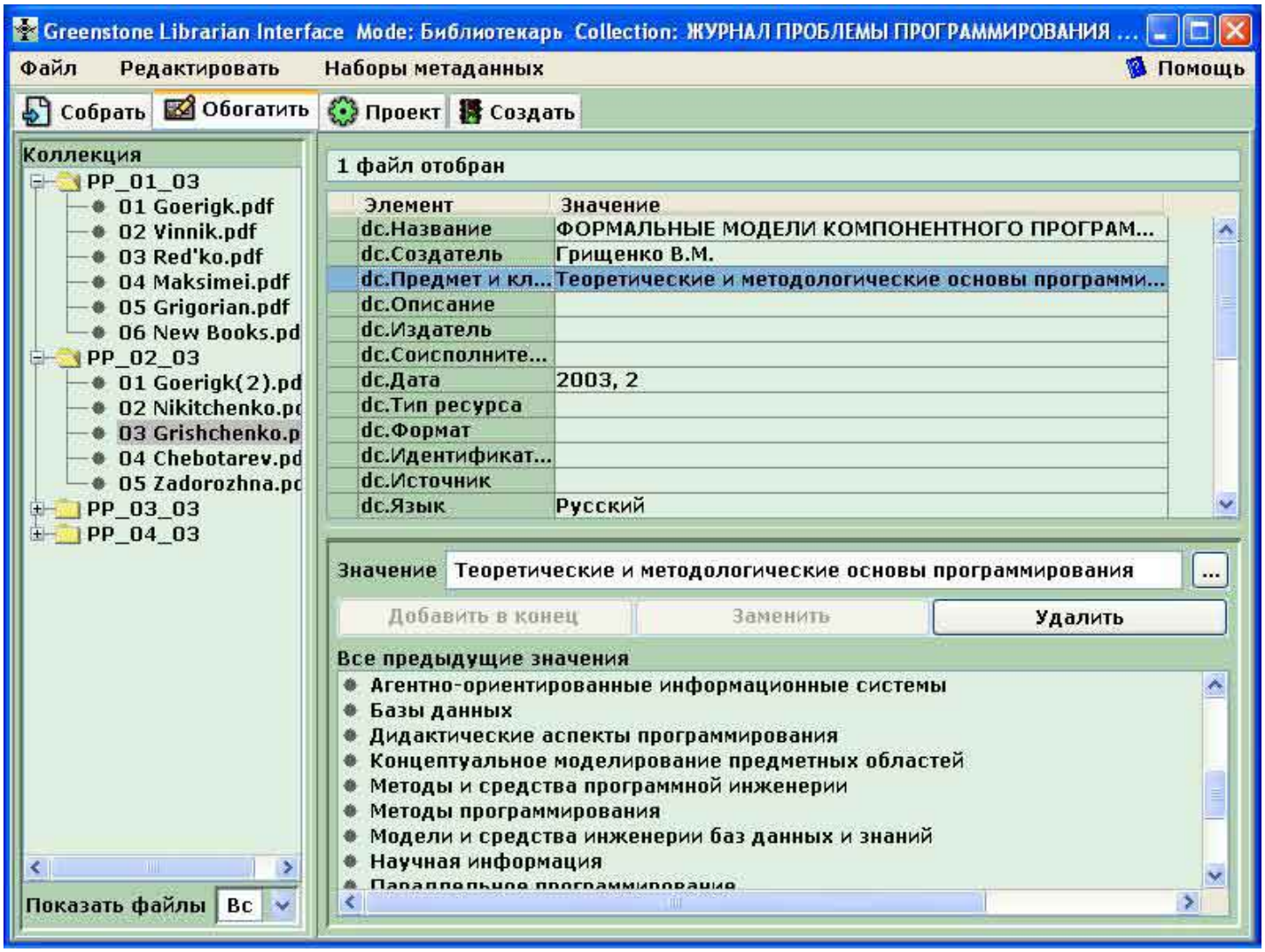

Рис. 14 Назначение метаданных документам коллекции

4.2.1 Сбор информации

Вначале пользователи либо открывают существующую коллекцию, либо начинают новую. На Рис.11 мы видим процесс запуска новой коллекции. Здесь вносится общая информация о коллекции — название и краткое описание содержания. В нашем примере мы связываем с коллекцией периодическое издание, поэтому на этом этапе мы будем использовать некоторые метаданные уровня периодического издания. Название - короткая фраза, используемая для идентификации коллекции в ЦБ (например, Журнал ПРОБЛЕМЫ ПРОГРАММИРОВАНИЯ). Описание - утверждение о принципах назначения коллекции и появляется под заголовком «*Об этой коллекции*» на домашней странице коллекции.

Далее пользователь решает строить коллекцию на основе существующей (Рис. 11) или проектировать новую. И если это новая коллекция, то для нее нужно выбрать один или несколько наборов метаданных (например, Дублинское Ядро [10]).

Кроме локального файлового пространства источником документов для коллекции может быть Web.

К новой коллекции могут добавляться документы, которые уже введены в другие коллекции ЦБ, но при этом могут возникать конфликты, поскольку их метаданные, возможно, входили в разные наборы метаданных, и GLI помогает пользователю разрешить эту ситуацию.

На Рис.12 показано интерактивное дерево файлов, используемое для просмотра локальной файловой системы. На данном этапе коллекция справа пуста, пользователь заполняет ее, перемещая нужные файлы с левой части экрана.

В системе предусмотрены специальные механизмы (фильтры) для работы с большими наборами файлов (Рис.13).

4.2.2 Добавление метаданных к документам

На следующем шаге построения коллекции нужно обогатить документы, добавив к ним метаданные. Это – то место, где пользователи GLI проводят большую часть своего времени: обогащение коллекций осуществляется с помощью выбора отдельных документов и ручного добавления метаданных (Рис.14). Мы уже упоминали две особенности GLI, помогающие справиться с этим заданием. Первая, документы, копируемые на первом шаге, прибывают с какими-либо подходящими присоединенными метаданными. Вторая особенность, всякий раз, когда это возможно, метаданные автоматически извлекаются из документов.

Для ускорения ручного ввода метаданных значения метаданных можно назначать нескольким документам сразу, на основании их общего пребывания в папке или посредством множественного выбора.

Назначенные ранее значения метаданных сохраняются и их легко использовать многократно.

В нашем примере мы ограничимся вводом следующих атрибутов документа типа публикация, которые покрываются стандартным набором

метаданных DC: Название; Автор; Предмет и ключевые слова; Дата; Организация; Язык.

Для полной реализации информационной модели (см. п. 2) метаданных простого DC недостаточно, в дальнейшем для более детального описания предполагается использовать расширенное DC с уточнителями элементов и с различными классификационными схемами.

В системе предусмотрена возможность просмотра всех метаданных, назначенных коллекции (Рис.15).

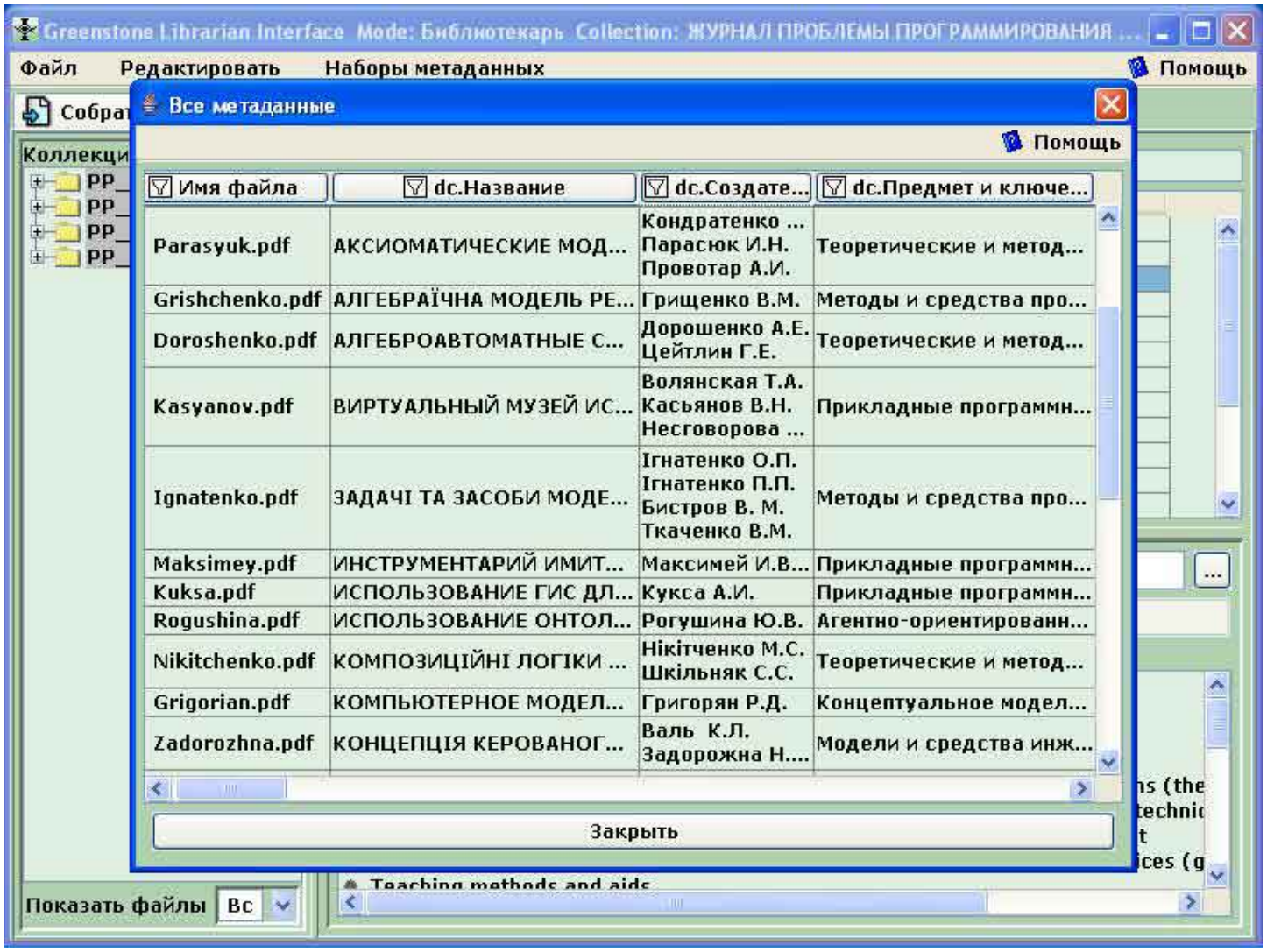

Рис. 15 Просмотр всех метаданных, описывающих выбранные документы.

4.2.3 Проектирование коллекций

Когда к файлам добавлены нужные метаданные, далее нужно решить, в каком виде коллекция Greenstone должна быть представлена пользователям. Это настраивается на этапе проектирования, и включает: общую информацию о коллекции, плагины документа, типы поиска, индексы поиска, индексы разбиения, поиск в нескольких коллекциях, просмотр классификаторов, элементы форматирования, перевод текста и наборы метаданных. Результат этого процесса регистрируется в файле конфигурации коллекции [5].

На этом этапе рецензируются и редактируются метаданные уровня коллекции, например, заголовок, автор и общедоступность коллекции. Есть возможность определить, какие должны быть построены полнотекстовые индексы. Можно создавать подколлекции. Можно добавлять или удалять поддержку определенных языков интерфейса. Здесь нужно решить, какие будут включены форматы документов (с помощью плагинов). Каждому плагину может понадобиться конфигурирование, определяемое соответствующими аргументами. Проектировщик коллекции должен определить, какие в Greenstone будут созданы структуры просмотра, они формируются модулями, называемыми “классификаторами”, которые также имеют разные аргументы. Также необходимо определить форматирование разных пунктов (элементов документа) в интерфейсе пользователя коллекции. Для всех этих элементов существуют обычно применяемые значения по умолчанию.

Поскольку Greenstone постоянно развивается и является системой Open Source, по мере того как разработчики добавляют в систему новые возможности, число опций увеличивается. Для того чтобы справиться с этим, Greenstone имеет сервисную программу ”информация плагинов“, которая вносит в список опции, доступные для каждого плагина, и GLI автоматически вызывает этот список, чтобы определить, какие опции показывать.

На Рис.16-20 дана иллюстрация выполнения процесса проектирования. Пользователь может выбрать отдельный интерактивный экран, каждый из которых связан с одним аспектом проектирования коллекции. Фактически он служит графическим эквивалентом ручного процесса редактирования файла конфигурации коллекции.

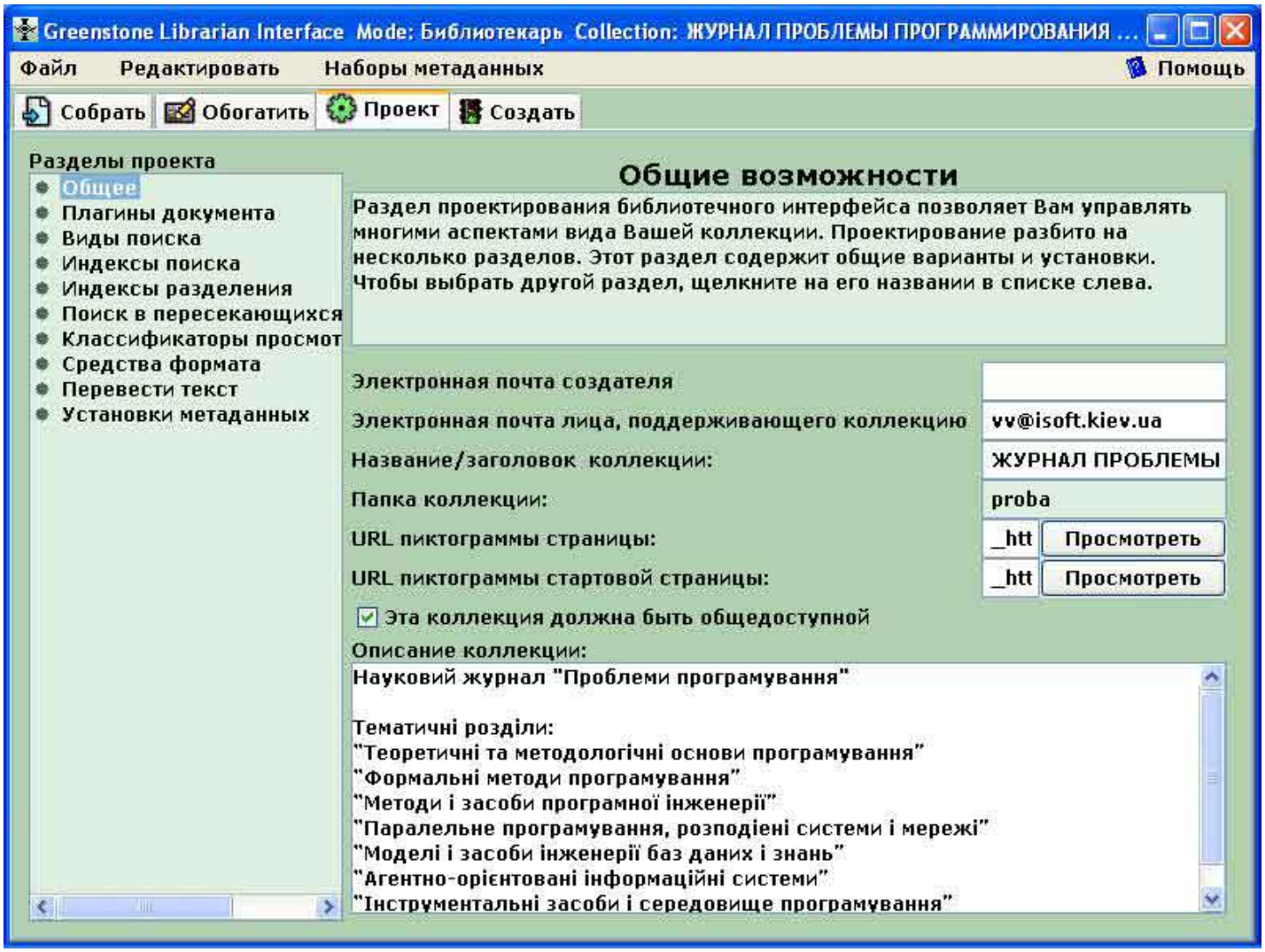

Рис. 16 Проектирование коллекции

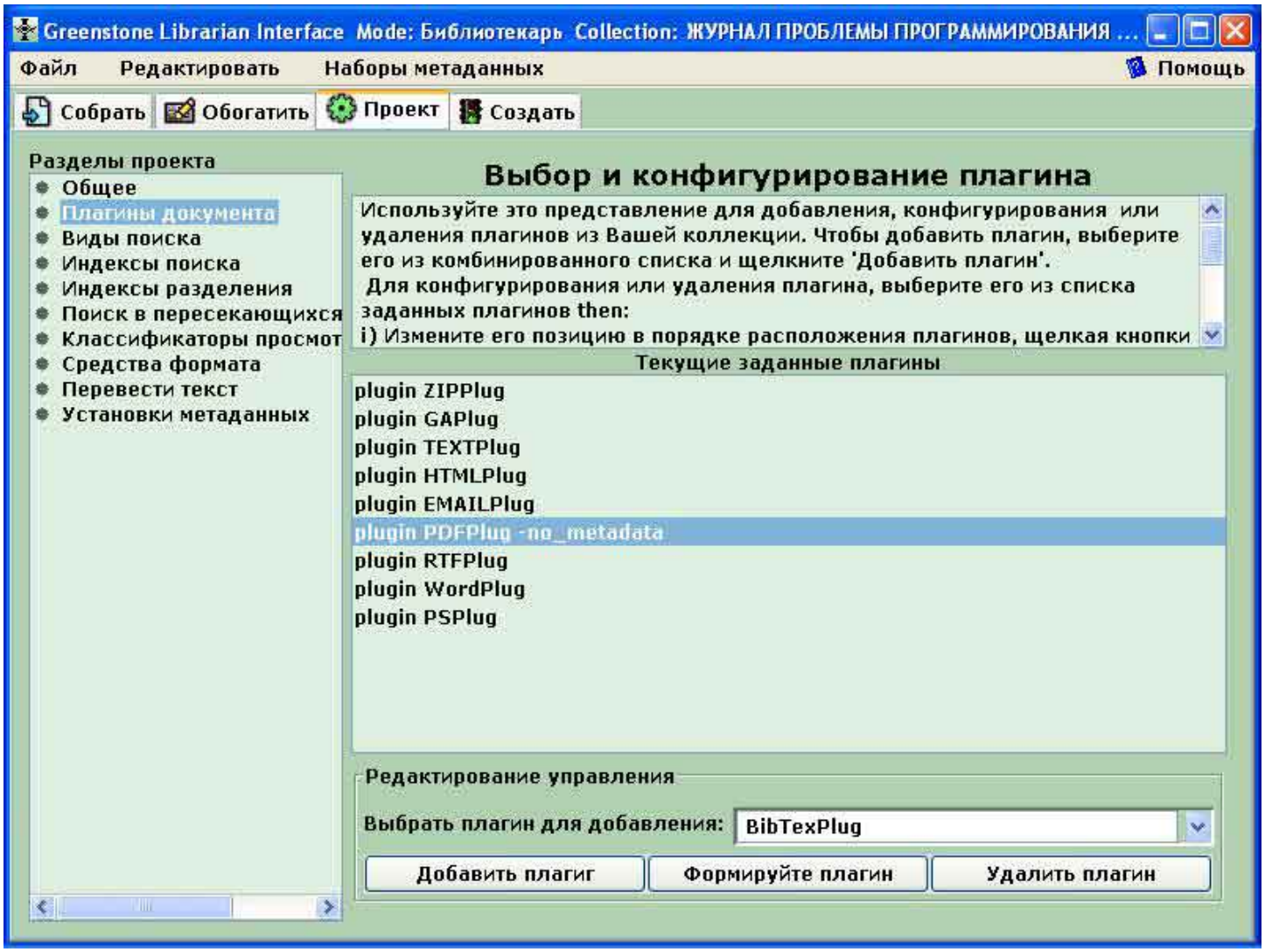

Рис. 17 Определение используемых плагинов

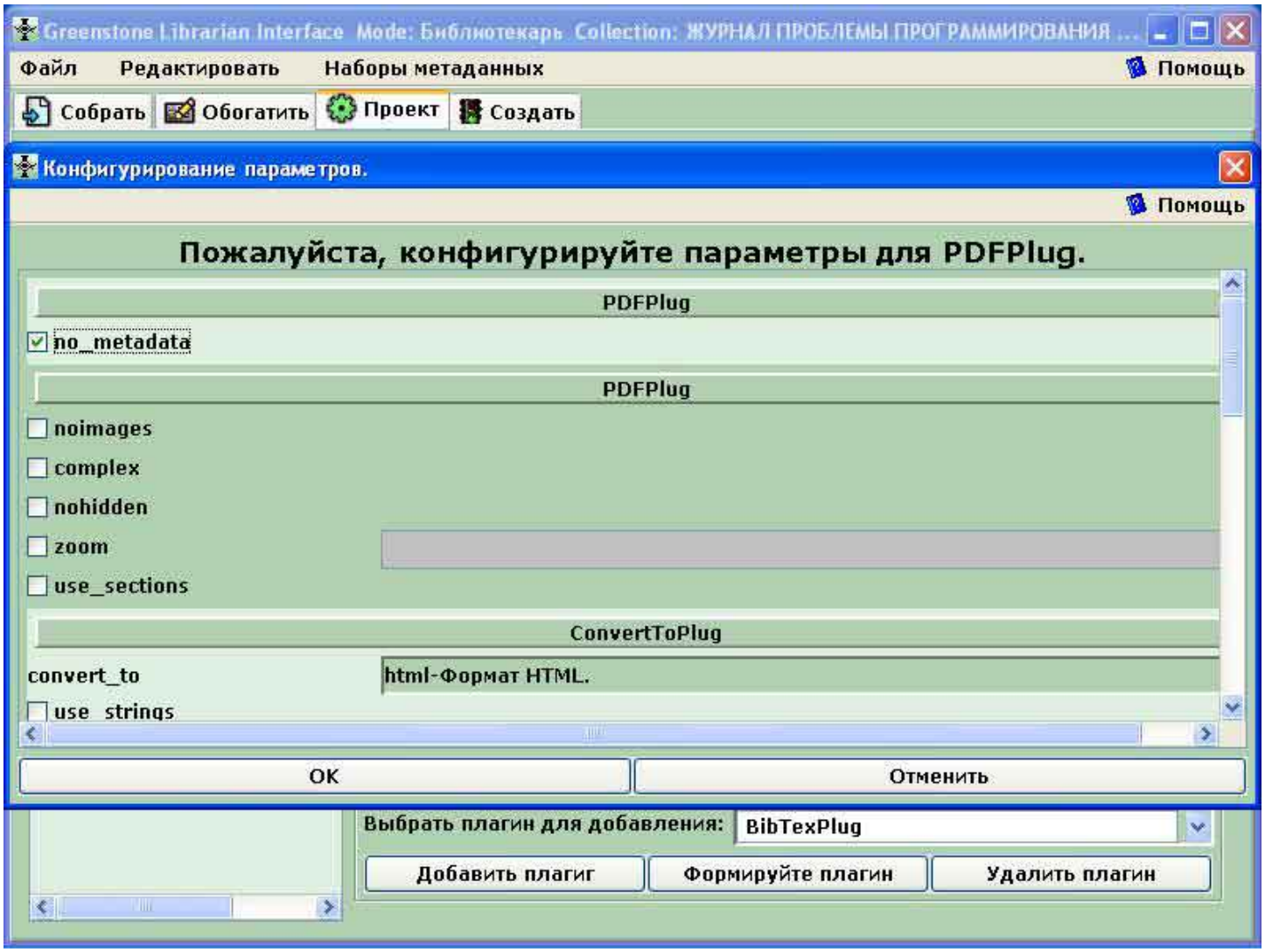

Рис. 18 Конфигурирование аргументов плагина

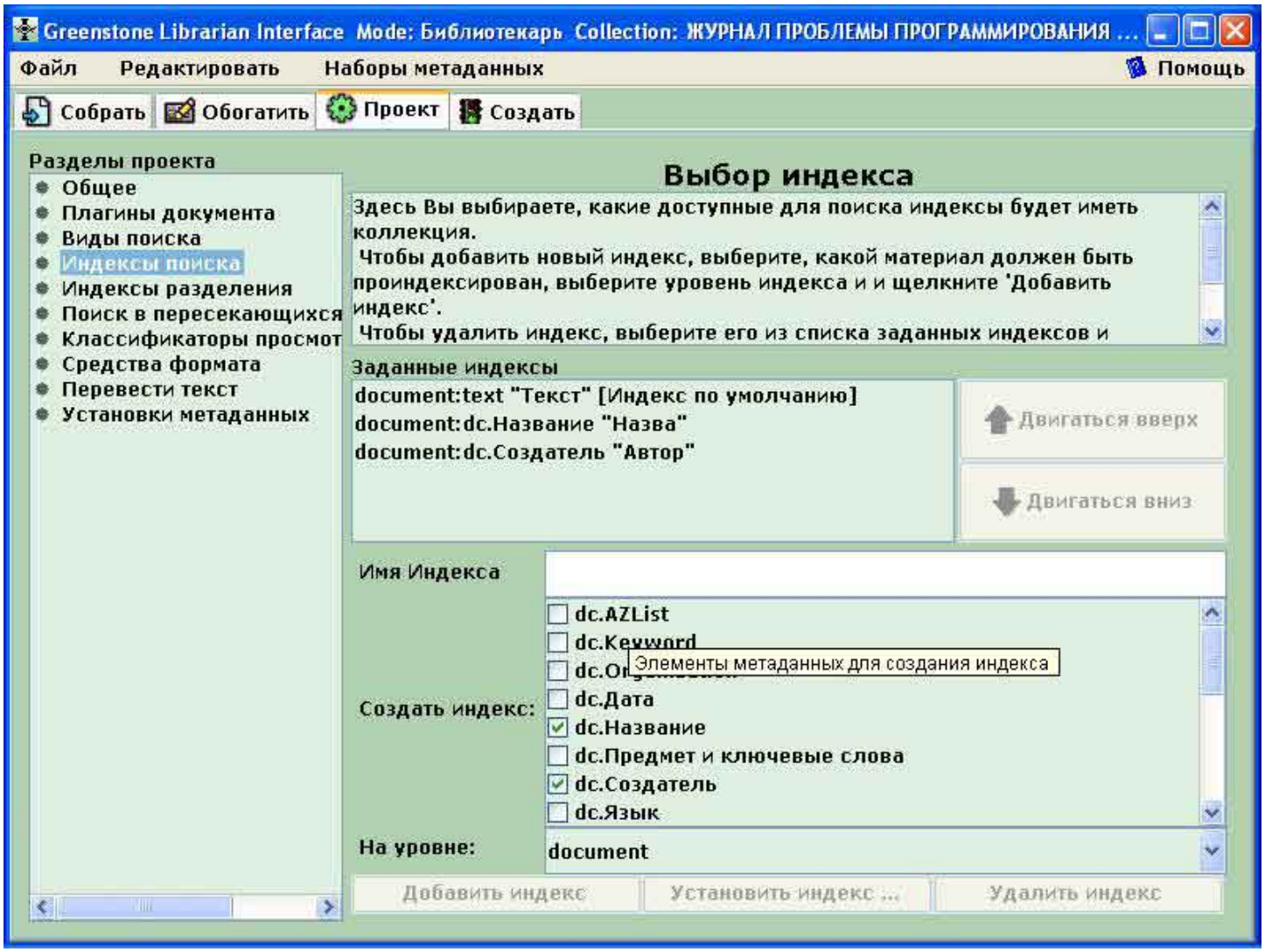

Рис. 19 Добавление индексов полнотекстового поиска

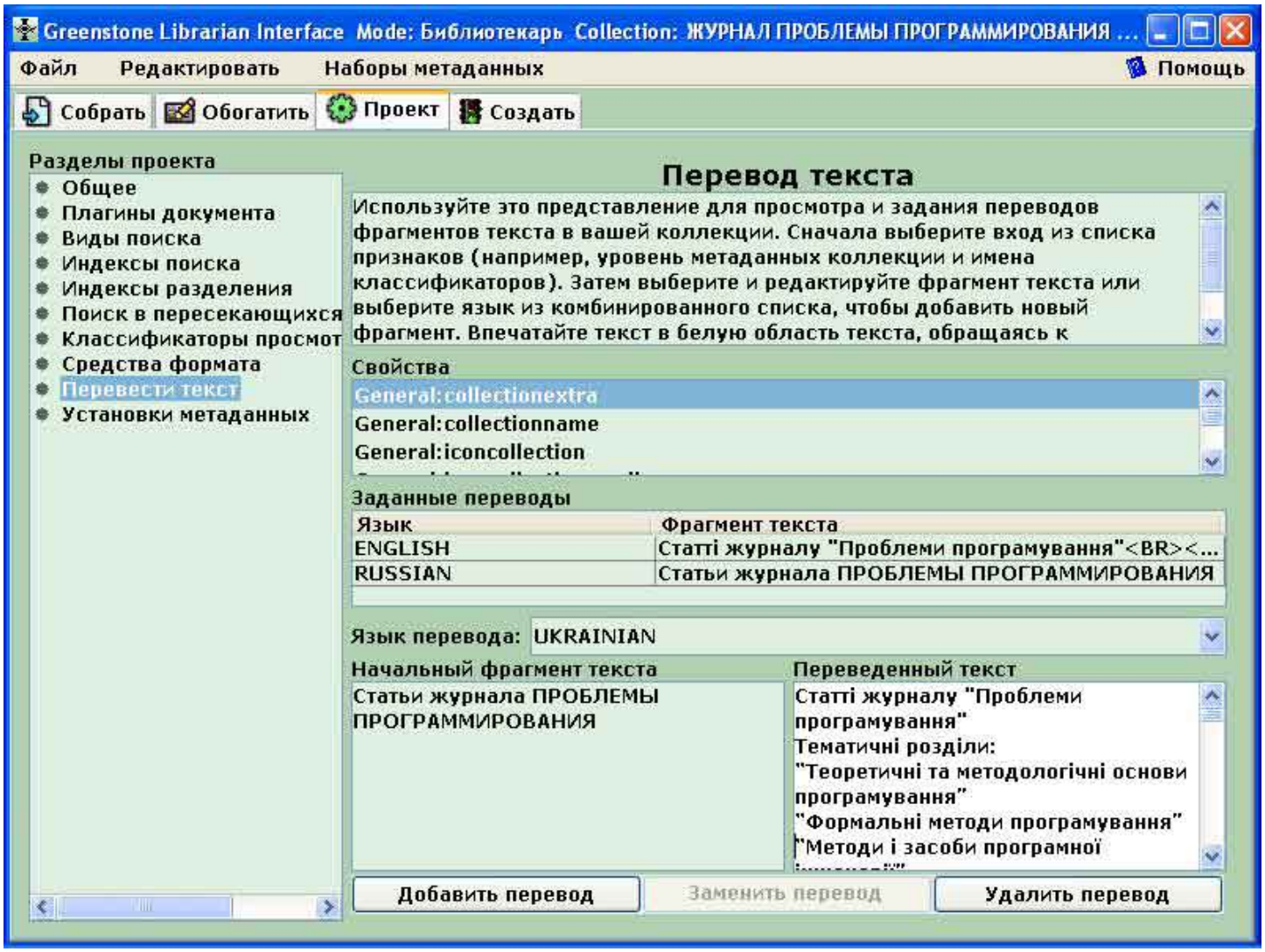

Рис. 20 Перевод фрагментов текста коллекции

4.2.4 Построение колекции

Следующий шаг - построение коллекции, сформированной из документов и описывающих их метаданных. Если построение завершается успешно, пользователи имеют возможность сразу выполнить просмотр созданной коллекции (Рис.1-10).

Таким образом, мы описали, как можно создать и настроить коллекцию документов, используя ПО Greenstone и интерфейс библиотекаря GLI. Мы показали, как можно реализовать трехуровневую модель информационного ресурса: научное издание - выпуск – публикация. Научное издание в нашей системе представлено коллекцией, и весь или частичный набор атрибутов научного издания приписывается ей. Выпуск для нашего примера представляется входной папкой коллекции, где собраны все документы с одинаковой датой. Наконец, публикация - это документ коллекции, и атрибуты публикации приписываются документу.

Систему можно доработать, используя DC-Library Application Profile [12]. В данном примере недостаточно решена проблема гетерогенности описаний ресурсов, например, применение разных схем предметных классификаций: "DC.Subject" scheme="ББК" и "DC.Subject" scheme="УДК". Для этого требуется подключение уже опубликованных отображений (mapping), либо реализацию новых алгоритмов отображения и выравнивания, как для наборов метаданных, так и для некоторых их атрибутов.

Заметный недостаток системы – постоянная природа индексных файлов, генерируемых в процессе построения, увеличивает стоимость модификации коллекции. Другой недостаток – низкий уровень функциональности, поддерживаемый системой Greenstone во время выполнения, хотя незначительные изменения легко поддерживаются, более существенные изменения включают модификацию и перекомпиляцию исходного кода. Далее вкратце рассмотрим новый проект Greenstone 3, который обещает преодолеть эти недостатки [13].

5 Проект Greenstone 3 как перспективный инструмент реализации ЦБ

Greenstone 3 нацелен на улучшение динамической природы инструментария по организации содержания и обеспечению сервисами при одновременном снижении накладных расходов, которые несут разработчики коллекций для достижения такой гибкости. Проект основан на современных стандартах, таких как стандарты платформы XML (в частности, XSLT, языка преобразования структуры XML-документа); современных реализациях методологий сообщающихся агентов (communicating agents); современных технологиях ПО, например, протоколах SOAP (Simple Object Access Protocol); современной стратегии кросс-платформенной разработки; современных схемах для модульного и динамического обновления ПО. Наиболее важно, что новый проект учитывает опыт предыдущих версий Greenstone, проблемы и трудные вопросы, с которыми сталкивались реальные пользователи, реальные разработчики коллекций, реальные библиотекари.

Далее кратко остановимся на некоторых важных характеристиках новой системы, которые существенно улучшат ее свойства.

Новая разработка имеет обратную совместимость, что дает дополнительные преимущества и обеспечивает разработчиков и пользователей легким способом миграции.

Для облегчения работы есть разные пользовательские уровни разных категорий персонала, которые принимают участие в построении ЦБ, например, разработчики содержания, редакторы коллекций, проектировщики последовательности выполняемых действий, разработчики ПО.

Модульный принцип организации кода - основа любого программно-инженерного подхода. Этому способствует применение существующих технологий: систем БД, инструментов индексирования, ПО визуализации страницы, использование стандартов.

Другой путь реализации модульности - построение ЦБ на наборе сервисов, в этом случае, модульного принципа функционирования.

Богатая инфраструктура цифровой библиотеки поддерживается распределенной архитектурой и открытым протоколом для интероперабельности. Запуск ЦБ на одной машине - тривиальный случай.

Старые коллекции можно представить в будущих версиях системы.

Многие аспекты библиотеки являются динамическими. Это охватывает и динамическое содержание, когда могут добавляться документы и метаданные, которые изменяются и удаляются, когда репозитарий находится в оперативном режиме и динамическую конфигурацию, что позволяет упорядочить вопросы презентации и добавления сервисов во время выполнения.

ПО использует систему интегрированной документации и т.д.

6 Заключение

ЦБ можно рассматривать как организованные, специализированные коллекции информации. Они сконцентрированы на отдельном предмете или теме, и хорошие цифровые библиотеки хорошо разъясняют принципы управления тем, что они содержат. Они создаются для того, чтобы информация стала доступной, четко определенной, и будут включать описание того, как она организована.

Актуальное значение имеет полноценное использование возможностей перспективных информационных технологий для практической реализации ЦБ. Хорошо известная в мире система Greenstone интегрирует современные технологии, поэтому ее использование будет полезно как научным работникам, так и разработчикам и пользователям ЦБ.

В данной работе подытоживается некоторый опыт работы с данной системой, сделан обзор современного ПО Greenstone, а также перспектив его развития. Построена модель информационного ресурса, имеющего трехуровневую структуру: периодическое издание, выпуск, публикация. На примере одного журнала описана процедура создания коллекции ЦБ с помощью интерфейса библиотекаря, входящего в состав ПО Greenstone.

Литература

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.-Петербург, 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. *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.
9. *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.
10. http://dublincore.org/usage/terms/dc/current elements/ Using Dublin Core - The Elements.
11. Adoption of Dublin Core by Governments/ http://dublincore.org/news/adoption/
12. DC-Library Application Profile (DC-Lib) - http://dublincore.org/documents/library-application-profile/
13. *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

---

Об авторах

Резниченко Валерий Анатольевич - старший научный сотрудник Института программных систем НАН Украины, г.Киев,
E-mail: reznich@isofts.kiev.ua

Проскудина Галина Юрьевна - научный сотрудник Института программных систем НАН Украины, г. Киев,
E-mail: gupros@isofts.kiev.ua

Овдей Ольга Михайловна - аспирант Института программных систем НАН Украины, г. Киев,
E-mail: olga_ovdiy@ukr.net
