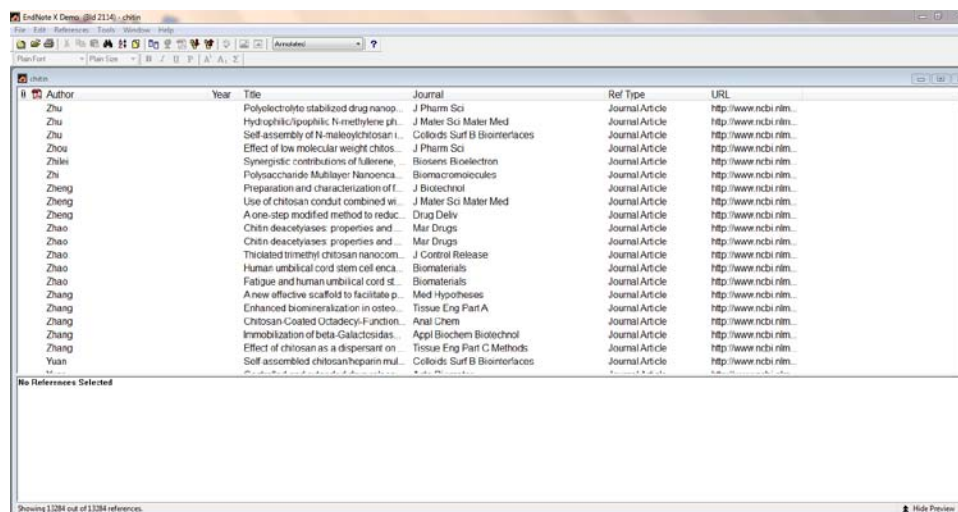


Техзадание для создания библиографической базы данных с возможностью поиска по полным текстам

База должна быть с открытыми кодами

Поля базы данных – автор, название, альтернативное название, ключевые слова, абстракт, номер патента, источник, год, том, номер, страницы, место издания, издательство, тип публикации, язык публикации, время добавления записи в базу, место хранения полнотекстовой публикации.

Просмотр записей и интерфейс главной странички – удобнее всего представить так, как это сделано в Endnote. И вполне реально для выполнения.



The screenshot shows the EndNote X6 Demo application window. The main area displays a list of references in a table format. The table has six columns: Author, Year, Title, Journal, Ref Type, and URL. The data is as follows:

Author	Year	Title	Journal	Ref Type	URL
Zhu		Polyelectrolyte stabilized drug nanop...	J Pharm Sci	Journal Article	http://www.ncbi.nlm...
Zhu		Hydrophilic/lipophilic N-methylene ph...	J Mater Sci Mater Med	Journal Article	http://www.ncbi.nlm...
Zhu		Self-assembly of N-maleoylchitosan i...	Colloids Surf B Biointerfaces	Journal Article	http://www.ncbi.nlm...
Zhou		Effect of low molecular weight chitos...	J Pharm Sci	Journal Article	http://www.ncbi.nlm...
Zhu		Synergistic contributions of fullerene...	Biosens Bioelectron	Journal Article	http://www.ncbi.nlm...
Zhi		Polysaccharide Multilayer: Nanosens...	Biomacromolecules	Journal Article	http://www.ncbi.nlm...
Zheng		Preparation and characterization of f...	J Biotechnol	Journal Article	http://www.ncbi.nlm...
Zheng		Use of chitosan conduit combined wi...	J Mater Sci Mater Med	Journal Article	http://www.ncbi.nlm...
Zheng		A one-step modified method to reduc...	Drug Deliv	Journal Article	http://www.ncbi.nlm...
Zhao		Chitin deacetylases: properties and...	Mar Drugs	Journal Article	http://www.ncbi.nlm...
Zhao		Chitin deacetylases: properties and...	Mar Drugs	Journal Article	http://www.ncbi.nlm...
Zhao		Triclated trimethyl chitosan nanocom...	J Control Release	Journal Article	http://www.ncbi.nlm...
Zhao		Human umbilical cord stem cell enca...	Biomaterials	Journal Article	http://www.ncbi.nlm...
Zhao		Fatigue and human umbilical cord st...	Biomaterials	Journal Article	http://www.ncbi.nlm...
Zhang		A new effective scaffold to facilitate p...	Med Hypotheses	Journal Article	http://www.ncbi.nlm...
Zhang		Enhanced biomineralization in osteo...	Tissue Eng Part A	Journal Article	http://www.ncbi.nlm...
Zhang		Chitosan-Coated Octadecyl-Functio...	Anal Chem	Journal Article	http://www.ncbi.nlm...
Zhang		Immobilization of beta-Galactosidas...	Appl Biochem Biotechnol	Journal Article	http://www.ncbi.nlm...
Zhang		Effect of chitosan as a dispersant on...	Tissue Eng Part C Methods	Journal Article	http://www.ncbi.nlm...
Yuan		Self-assembled chitosan/heparin mul...	Colloids Surf B Biointerfaces	Journal Article	http://www.ncbi.nlm...

At the bottom of the window, it says "Showing 1284 out of 1324 references." and there is a "Hide Preview" button.

Таблица, где по умолчанию находятся 6 столбцов. Каждая строка – это одна публикация.

Столбцы.

Первый – кнопка с иконкой pdf файла. Если есть – есть полнотекстовая публикация. Если «нет» - понятно - отсутствует. Щелчок по кнопке – переход к полнотекстовой публикации. Щелчок по пустому месту или другой иконке на этом месте позволяет присоединить полнотекстовую публикацию.

Второй – Автор(ы). Щелчок по верхней строке столбца правой кнопкой мышки приводит к сортировке по возрастанию. Повторный щелчок – по убыванию.

Вторая столбец – год издания публикации и аналогичные возможности с сортировкой.

Третий столбец – название (и сортировка)

Четвертый столбец – источник (и сортировка)

Пятый столбец – тип публикации (и сортировка)

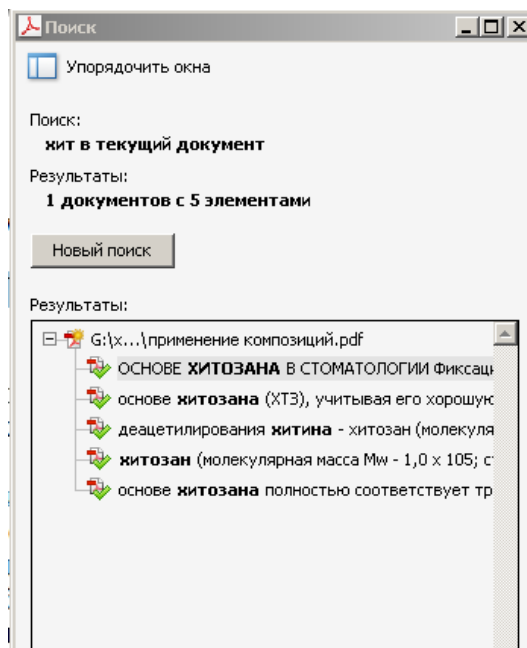
Внизу находится окошко (которое условно можно назвать - реферат), в котором можно просмотреть более подробную информацию о публикации. Несколько вариантов просмотра выбираются в командном меню из ниспадающего списка.

В нашем случае есть смысл предусмотреть 4 варианта просмотра.

1 вариант (стандартный – автор, год, источник, том, номер, страницы, реферат – если патент или авторская заявка – то вместо столбцов (источник, том, номер) - номер патента или заявки)

2 вариант – показать все. Вся информация из полей библиографии для данной публикации.

3 вариант. Показать результат поиска. В этом случае записи найденных публикаций выводятся как и в первом примере, но расположены по релевантности, но в окошке «реферат» выводится описание предмета поиска (например «деацетилирование хитина») с указанием количества найденных совпадений для выбранной из списка (стрелками или мышкой) публикации и все варианты совпадений в контексте – два слова слева от найденного слова и десять слов справа, размещенные списком (как в поиске Acrobat)



И четвертый вариант – «выбрать другой стиль». Это на всякий случай – когда-нибудь на эту кнопку можно будет добавить настраиваемые возможности форматирования ссылок в соответствии с требованиями ВАК и разных журналов.

В самом низу панели располагаются. Слева – информация об общем количестве записей в библиотеке. Справа – кнопка для удаления нижнего окошка с рефератом и расширения, посредством этой операции, списка записей.

Разделы верхнего ниспадающего меню.

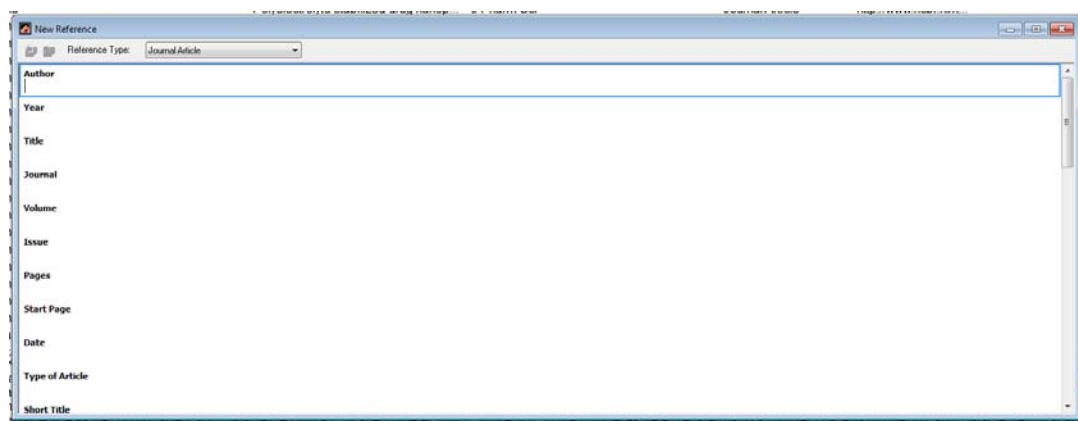
«Файл», «правка», «записи» «инструменты», «вид», «?», «информация о базе данных»

Меню «Файл» включает команды «создать библиотеку», «открыть библиотеку», «сохранить библиотеку», «сохранить как», «удалить библиотеку», «записать библиотеку в...», «объединить библиотеки», «записать библиотеку из...», «печать», «настройки печати», «предварительный просмотр», «выход»

Меню «Редактирование» - включает стандартные команды - «отменить ввод», «повторить ввод», «выделить все», «копировать», «вырезать», «вставить», «изменить текст», «фонт», «размер», «стиль»

Меню «Записи» включает команды – «новая запись», «редактировать записи», «сохранить записи», «удалить записи», «найти записи», «следующая запись», «предыдущая запись», «показать все записи», «показать выбранные записи», «скрыть выбранные записи», «сортировать записи», «скрыть записи», «найти дубли», «добавить PDF файл к записи». Еще возможна команда «получение записей из внешней библиотеки», но это серьезная работа. Потом ее можно будет добавить в виде отдельного модуля. Сейчас об этом думать рано.

Создание новой записи удобно производить через использование специальной формы, которая открывается командой «новая запись» и дублирует аналогичный процесс в «Endnote». Для данной цели создается таблица с одним столбцом. Каждая строка содержит неизменяемое название и пустое поле для ввода соответствующей информации. Перемещаться по столбцу можно при помощи мышки и клавиши «таб». Желательно, чтобы поля были специализированными для различного типа данных, с тем, чтобы ввести текстовую информацию в поле для цифр было нельзя. В полях «автор», «журнал», «ключевое слово» при наборе текста включается режим «подсказка» из автоматически создающихся при введении записей библиотек (об этом ниже). Так при введении буквы «М» - выскакивает «МАН», после введения второй буквы «МУ» - предлагается «МУАН», после МУЗ «Muzzarelli», что соответствует нашему желанию и экономит несколько секунд (и позволяет избежать грамматических ошибок).

The image shows a screenshot of the 'New Reference' dialog box in the EndNote software. At the top, there is a dropdown menu for 'Reference Type' which is currently set to 'Journal Article'. Below this, there is a list of fields for entering bibliographic information. The fields are: Author, Year, Title, Journal, Volume, Issue, Pages, Start Page, Date, Type of Article, and Short Title. Each field has a corresponding input area. The dialog box has a standard Windows-style title bar and window controls.

Кроме того, необходимо предусмотреть возможность добавления больших массивов информации из внешних баз данных. Ручной перенос данных для больших массивов неприемлем. Обычно эту проблему решают с использованием библиотек фильтров. Но, поскольку внешних источников и библиотек много, разумнее запрограммировать перенос данных из стандартной таблицы Word или Excel. А конвертацию текстовых массивов (все издательства бесплатно позволяют копировать библиографические данные и рефераты из своих баз данных в текстовом виде) в табличный формат можно легко проделывать для любого формата – макросами. Хотел бы отметить – гладкое решение данной задачи - чрезвычайно актуально и востребовано.

Такой же важной задачей (даже проблемой) является поиск дублей из меню «Записи». Если невозможно идентифицировать одинаковые публикации, база данных превращается в кучу мусора и возможность ее расширения и даже поддержания - невозможна. Кстати, я не встречал качественного решения этой проблемы ни в одной программе, где эта функция была задействована и прорекламирована. В том же «Endnote», как и у меня в «Библиографии» - даже при усовершенствованном поиске – одинаковые публикации, полученные из различных источников или введенные вручную – могут не идентифицироваться в качестве дублей. Попадают два пробела вместо одного, отсутствие пробела, еще какие то знаки препинания могут не совпадать, фамилия автора и инициалы также отличаются по написанию, может

отсутствовать или наличествовать какое то из библиографических полей, авторы могут называться редакторами или наоборот и т.д. **В любом случае эту проблему необходимо решить.**

Должен сработать следующий вариант (правда, не очень понятно, каким образом его реализовать). Я предлагаю поиск дублей производить по полю «название» и использовать два режима поиска. Первый поиск должен быть индексированным и искать нужно не точное соответствие, а максимальное совпадение. А затем первые из найденных трех-четырех (в числе наиболее совпадающих или релевантных) вариантов следует обрабатывать через функцию прямого поиска. Совпадение двух вариантов поиска – позволит автоматически сделать заключение о совпадении записей. При отсутствии совпадения – компьютер должен создать список и показать его на экране (вместе с текстом, который использовался для проведения поиска,) чтобы можно было оценить – есть дубли по названию в данной подборке или их нет.

Вероятно, я не понятно точно выразился.

Попробую объяснить на примере.

Название статьи «Методы определения степени деацетилирования β -хитина»

1 вариант поиска

Предложение разбивается на изолированный список (условно назовем его «ключевыми словами» и я привожу их в скобках) [(методы) и (определение) и (степень) и (деацетилирование) и (β -хитин)] - и ищет их сочетание в названиях статей, вычисляя коэффициент совпадения. Желательно отбрасывать союзы, служебные слова и артикли и учитывать порядок слов. Я видел такие приложения Active X.

Фантазируем. Пусть наиболее релевантными из найденного оказываются:

«Методы определения степени деацетилирования бета-хитина»

«Методы исследования степени деацетилирования β -хитина»

«Современные методы определения степени деацетилирования»

«Методы определения степени деацетилирования хитина»

Запускается прямой поиск по точному совпадению в этих предложениях и ничего не находится (если демонстрируется полное совпадение – это дубли) и информация об этом выводится на экран в виде стандартных строчек и возможностью удалить повторяющиеся публикации.

Когда прямой поиск не находит полного совпадения шанс на присутствие дубля все-таки остается. В описываемой ситуации на экран выводится список в виде стандартной строчки, который содержит все пять публикаций (pdf файл, год, автор, название, источник)

1. «Методы определения степени деацетилирования β -хитина», Петров, 2010

2. «Методы определения степени деацетилирования бета-хитина», Петров, 2010

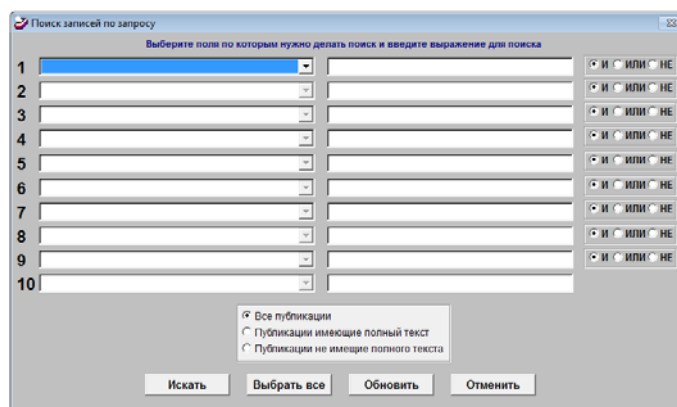
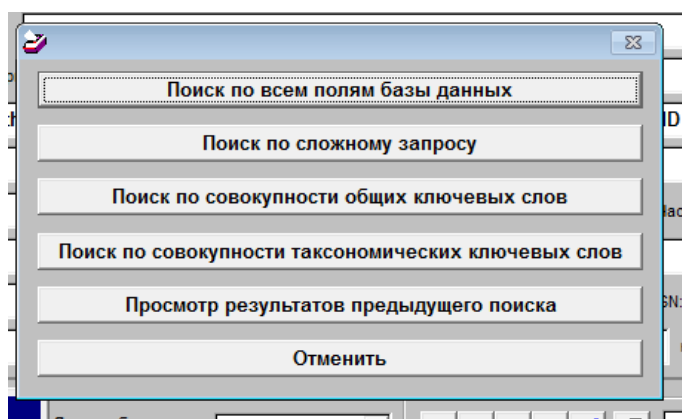
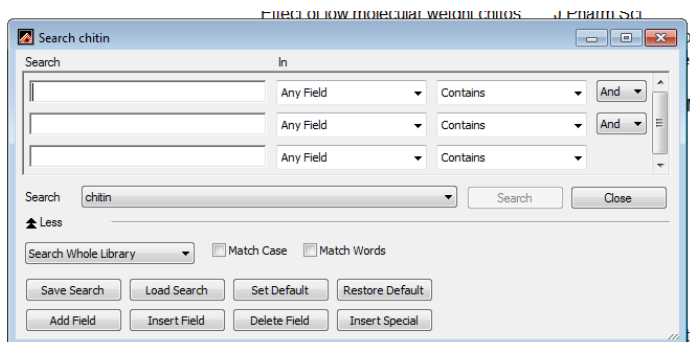
3 «Методы исследования степени деацетилирования β -хитина», Варламов, 2009

4 «Современные методы определения степени деацетилирования», Ильина, 1996

5 «Методы определения степени дезацетилирования хитинов», Большов, 2010

При просмотре этого списка понятно, что скорее всего 1-ое и второе предложение являются дублями, а ошибка вылезла из-за невозможности использовать специальные символы в тексте.

Одна из самых важных команд «Поиск записи» Здесь также очень удобно использовать форму, принятую в «Endnote» Она выглядит следующим образом. В целом, похоже на поиск из «библиографии», но первый вариант, похоже, удобнее. В «Endnote» есть очень удобные функции – «вставить поле», «убрать поле», «добавить поле», что позволяет придавать поиску нужную степень глубины. Применение данных функций либо добавляет еще один ряд с условиями поиска (выбор полей, в которых проводится поиск, условия поиска («содержит», «точное соответствие», «больше или совпадает», «меньше или совпадает», «начинается с», «заканчивается на») и оператор) - либо удаляет выбранный ряд.

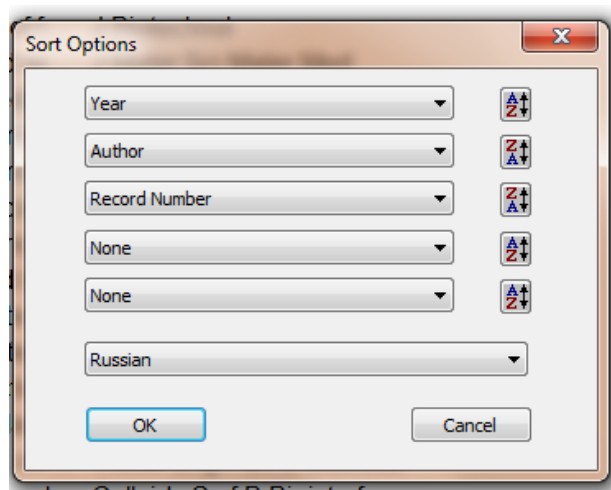
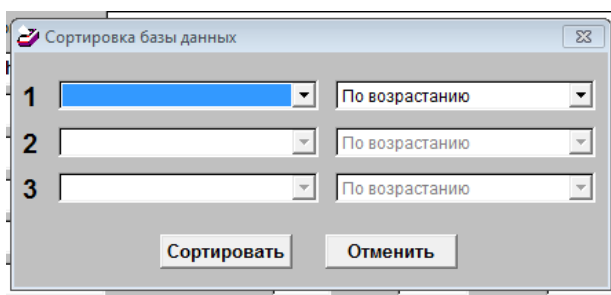


В целом, можно использовать форму «Endnote» но либо к полям для поиска добавить еще одно поле «поиск по полным текстам», либо предусмотреть запуск данной функции прямо на интерфейсе формы, включая нужный флажок (как на вышеприведенном примере с выбором

полнотекстовых либо неполнотекстовых публикаций). Кроме того, следует предусмотреть выбор способа поиска. Индексированный или прямой. Несмотря на относительную медлительность – прямой поиск обладает и некоторыми преимуществами (например, может искать части слов). Функцию «выбор типа поиска» можно разместить в каждом ряду (перед или позади функции «выбор оператора»), но, вероятно, можно разместить и в другом месте.

Еще одна крайне полезная для пользователя функция – возможность проверки наличия публикации в базе данных без открытия этого приложения (удобная, упрощенная и ускоренная модификация команды «Поиск записи».) Очень часто в интернете или при чтении электронной публикации в тексте появляется ссылка на публикацию и необходимо проверить, присутствует ли данная публикация в базе данных или ее следует скачать (в основном моя библиотека пополняется таким образом). Подобная функция в Lingvo реализована следующим образом. В тексте выделяют слово или словосочетание, а затем нажатием клавиш (как правило ctrl + ins + ins) активируется вызов программы и осуществляется перевод. Вы можете возразить, что тут не нужна спешка, можно выйти на рабочий стол, открыть программу, запустить команду «поиск по названию» или «поиск по автору», перейти в интернет, скопировать нужный термин, вернуться в «библиографию», вставить термин в нужное поле, запустить поиск и посмотреть – есть ли это название, если ничего не найдено, вернуться в интернет, скопировать автора, вернуться в программу, отыскать все публикации данного автора, просмотреть их и уже точно определить – наличествует данная публикация в базе или ее нет. Это, конечно, так. Но перевод неизвестных терминов при помощи трех клавиш я делаю постоянно и сомневаюсь, что у меня возникло бы это желание при выполнении стандартных процедур запуска программы и копирования текста в поле перевода.

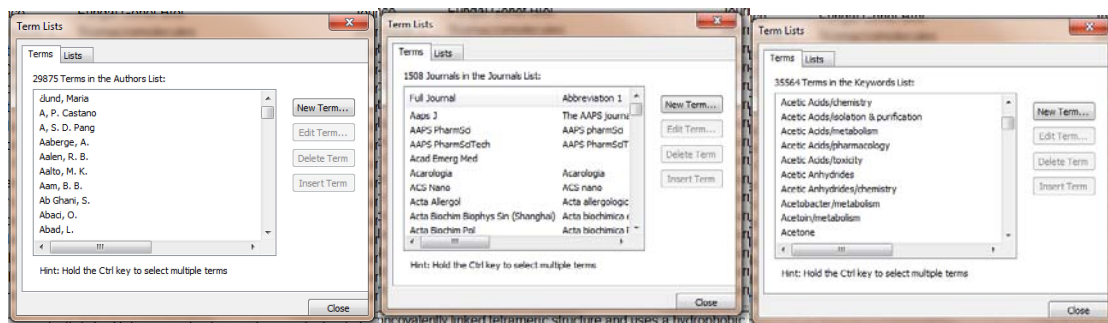
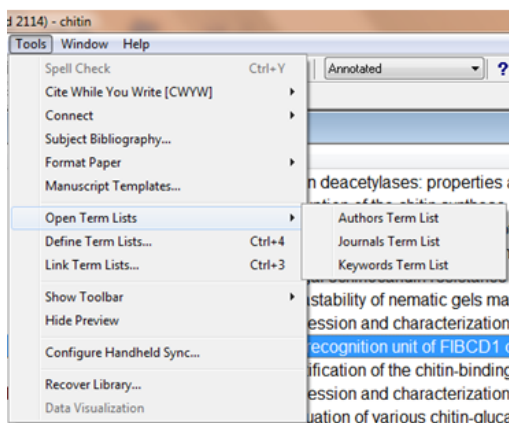
Команда «сортировать записи» и в «библиографии» и в «Endnote» отличаются незначительно, хотя пожалуй, что второй вариант несколько предпочтительнее.



Ну и последняя команда из меню «Записи» - «добавить PDF файл к записи» не требует комментариев. После активации данной команды появляется дерево файлов, на котором можно выбрать маршрут. Хотя теперь я вижу, что более удобна ситуация, когда папка с полнотекстовыми публикациями находится в корневом каталоге папки «библиография», а команда «создать библиотеку» - автоматически создает новую папку – «библиотека 2» в том же месте.

Чтобы исправить всякие перекосы в библиографии следует при помощи скрипта автоматически переименовать все файлы в библиотеке (присвоив им порядковые номера и автоматически переименовав название файлов в маршрутах). Придуманные мной сложные названия на русском и английском языке для файлов себя не оправдали (хотя это было не просто так и логика этого процесса существовала). С такими именами происходят перекосы при работе программы на различных компьютерах. Более того – часть публикаций теряется при копировании. Это вы обнаружили на своем примере.

Команды в меню – «инструменты» - пока трудно точно сформулировать. Я надеюсь, что тут будут располагаться еще не написанные модули для интеллектуальной обработки текста, а также некоторые библиотеки. Я уже писал, что в «Endnote» работает механизм автоматического создания списков авторов, источников и ключевых слов, которые затем предоставляются в виде алфавитных списков. Эти списки можно просматривать, редактировать и копировать для различных нужд. Кроме того, списки используются также для создания новых записей, оформления поисковых запросов и т.д., как я описывал выше.



Мне кажется, что пользуясь этими тематическими библиотеками (авторов, источников и ключевых слов) можно создать исключительно удобный метод вызова всех библиографических записей, которые содержат нужный термин или их совокупность. Я не вижу сложностей для написания подобного модуля, разве что он окажется тяжеловат для процессора.

В этом же меню «Инструменты» следует разместить и «Менеджер соединений», но это только в том случае и после того, как будет написан модуль для связи с различными библиотеками. Сейчас я этот пункт (как и несколько вышеописанных) в ТЗ не включаю.

Меню «Вид» предназначено для управления различными способами расположения открытых окон и карточек на экране («каскад», «упорядочить все», «скрыть все», «масштаб»).

Меню «Help» в объяснении не нуждается.

И еще несколько пожеланий.

Необходимо, чтобы основные команды запускались при помощи сочетания клавиш

Стандартные команды выделения, копирования, вырезания и переноса выполнялись во всех окнах мышкой

Чтобы все сделанные изменения сохранялись лишь после запроса на подтверждение сохранения сделанных изменений.

Готовая программа с базой данных (на момент написания) должна существовать в виде свернутого пакета, который можно устанавливать на любом компьютере.