

Дунаев А.В.* , Корабейников И.Н.**
*ГОУ ВПО «Оренбургский государственный университет»,
**Оренбургский филиал ГУ «Институт экономики УрО РАН»,

**ПОИСКОВАЯ СИСТЕМА НАУЧНЫХ И ИССЛЕДОВАТЕЛЬСКИХ РАЗРАБОТОК
ИНСТИТУТОВ И ФИЛИАЛОВ ОРЕНБУРГСКОГО НАУЧНОГО ЦЕНТРА
УРАЛЬСКОГО ОТДЕЛЕНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК***

Представленная публикация посвящена итогам научно-исследовательской работы, проведенной при финансовой поддержке РГНФ. В статье представлены методические и научно-практические результаты создания поисковой интернет-системы разработок институтов и филиалов Оренбургского научного центра Уральского отделения Российской академии наук.

Лавинообразное накопление информации во внутренней сети и сети Интернет постоянно усложняет проблему нахождение необходимой информации. Общесистемные поисковые службы (<http://www.rambler.ru/>,<http://www.yandex.ru/>,<http://www.google.ru>,<http://www.aport.ru>,<http://www.search.ru>,<http://www.hotbot.com>,<http://www.lycos.com>,<http://www.yahoo.com>,<http://www.atrus.ru>,<http://www.deja.com>) плохо удовлетворяют требованиям ученых по систематизированному отбору данных из-за большого количества предоставляемых документов, которые находятся в неклассифицируемом состоянии. Для уточнения запросов необходимо хорошее знание возможностей конкретных поисковых программ таких поисковых систем как google.com, yandex.ru (точнее – специального языка поисковой машины), и, как следствие, на поиск нужной информации тратиться много времени.

Необходимо учесть то, что, как правило, простой запрос к существующей общетематической поисковой системе возвращает огромное число документов, которые никак не относятся к теме поиска. Этот факт резко снижает продуктивность не только поиска, но и работы человека из-за необходимости просмотра многочисленного информационного «мусора».

В настоящее время в сети Интернет функционируют сотни поисковых систем и тысячи коллекций документов, электронных каталогов, библиотек с различным сервисом и уровнем доступа (<http://promo.webcom.by/content.php?code=80>, <http://animapoisk.narod.ru>). Каждую из этих систем ведет администратор. Для публикации своих исследований в электронном каталоге ученый должен заполнять различные формы, в которых ему необходимо указать точную индексацию и предусмотреть все взаимосвязи сво-

ей работы с документами, находящимися в сети (в виде гиперссылок или в списке литературы).

Практически все электронные системы предусматривают дополнительные возможности по систематизации и поиску информации. Например, научная электронная поисковая система перспективных научно-исследовательских разработок <http://lib.ru> или зарубежный аналог <http://www.elis.sk>, <http://www.isinet.com>, в которых в основном информация систематизирована по авторам, отделам (организациям) и классификатору AMS. Система имеет внутренний поисковик и допускает развитие системы пользователями, путем публикации материалов в систематизированных разделах.

Вместе с тем все выше перечисленные системы имеют основные недостатки работы с информационно научно-исследовательскими разработками. Пользователь использует только внутреннюю базу системы, так как алгоритм поиска не предназначен для поиска по внешним источникам (библиотекам, каталогам, поисковым системам), этот недостаток есть как в наших системах, так и зарубежных, хотя наши проекты в большей степени предусмотрены для пользователей простотой и удобством использования. Поэтому необходим комплексный подход: в системе должны быть программные модули, обеспечивающие поиск, каталогизацию, взаимосвязь и конвертацию данных (для взаимодействия с различными библиотечными системами, поисковыми системами и каталогами) и возможность наращивания функций за счет надстроек над системой.

Информационные ресурсы, накопленные научно-исследовательскими учреждениями, находятся сейчас в разрозненном состоянии и, зачастую, оказываются недоступны их потребителю. Иногда, невозможно найти информа-

*Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (грант №05-02-81201 а/У)

цию даже о наличии той или иной разработки, имеющейся в хранилищах соседних учреждений. Поэтому есть необходимость в концентрации и доступности информационных ресурсов.

Работа посвящена решению актуальной проблемы, организации системы поиска информации научно-исследовательских разработок, регионального научного центра для целей повышения экономической эффективности деятельности Оренбургского научного центра УрО РАН (ОНЦ УрО РАН) путем формирования электронных поисковых коммуникаций как внутри научного центра, так и между научными центрами России и других стран мира.

Это является одной из первостепенных задач любого, объемного ресурса особенно научно-исследовательского центра. Созданная поисковая информационная система представляет собой качественное, доступное и удобное средство организации поиска по внутренней информационной базе или внешним страницам сайта без каких-либо вынужденных ограничений; средство, учитывающее особенности структуры и направления сетевого ресурса, его индивидуальные черты характера.

Проект направлен на создание информационно-поисковой системы состоящего из открытого и внутреннего электронного каталога и информационной базы знаний (каталоге информационных ресурсов сети Интернет) по проблемам экономики АПК, биотехнических систем, микробиологии, инфектологии, микробной экологии, геоэкологии, гидроэкологии, на создание системы сбора, каталогизации и поиска информации.

Проект разработан в виде интегрированной базы данных Оренбургского научного центра Уральского отделения Российской Академии Наук, который включает в себя информацию по разработкам Оренбургского филиала Института биотехнических систем УрО РАН, Оренбургского филиала Института клеточного и внутриклеточного симбиоза УрО РАН, Оренбургского филиала горного института УрО РАН, Оренбургского филиала института экономики УрО РАН. Поиск по информационной базе осуществляется с помощью специализированного программного обеспечения, поисковая система учитывает индивидуальность каждого научно-исследовательского направления. Так же система поиска взаимосвязана с внешними источниками (базами библиотек, с всемирными поисковыми машинами). Все это

в десятки, раз ускоряет поиск необходимой информации по сравнению с традиционными подходами. Работа выполнена при финансовой поддержке Российского гуманитарного научного фонда (грант №05-02-81201 а/У).

Поэтому есть необходимость в объединении информационных ресурсов в единое целое, что позволит рационализировать работу коллектива и повысить экономическую эффективность деятельности Оренбургского Научного Центра Уральского Отделения Российской Академии Наук.

В настоящее время перед региональными научными центрами возникла необходимость в качественном, доступном и удобном средстве организации поиска по внутренней информационной базе или внешним страницам сайта без каких-либо вынужденных ограничений; средство, учитывающем особенности структуры и направления сетевого ресурса ОНЦ УрО РАН, его индивидуальные «черты характера». Поэтому, проект направлен на решение актуальной проблемы – обоснованию и разработке системы поиска информации по внешней информационной среде (базы данных серверов всемирной сети Интернет) и внутренней базе данных регионального научного центра.

Круг пользователей не ограничен сотрудниками ОНЦ УрО РАН, пользователи системы – это все посетители интернет-сайта, кому необходимо проводить поиск и выборку информации по указанным инновационным направлениям. Ориентировочно, на начальной стадии работы системы, использовать ее будут около 2100 человек – научные сотрудники ОНЦ УрО РАН, докторанты других научных и учебных заведений города (ОГУ, ОГМА, ОГАУ, ОГПУ), работающие по указанным направлениям не считая пользователей интернет. После полного завершения работ будет обеспечена взаимосвязь со всеми основными центральными электронными библиотеками РФ по разным научным направлениям и число пользователей возрастет. Предусмотрено создание открытого интерфейса пользователя с целью обеспечения возможности подключения подобных систем других научных центров в единую базу знаний.

Существующие в сети опорные информационные узлы (ОИУ) и ассоциированные информационные узлы (АИУ), являющиеся поставщиками информации и обладающие соответствующими средствами подготовки, обработки и хранения документов. Опорные узлы

должны обладать развитой инфраструктурой, значительной вычислительной мощностью и информационной емкостью, иметь квалифицированный технический персонал. Немаловажным условием является возможность выхода в Internet через высокоскоростные и надежные каналы связи, такие как опто-волоконные. Основной задачей ОИУ является поддержка региональной информационной инфраструктуры с предоставлением сервисов различного типа: DB – сервиса, Mail – сервиса, News – сервиса, WWW – сервиса, Internet – сервиса для ассоциированных центров, Security – сервиса.

Ассоциированные узлы, представляющие основную массу поставщиков информации, могут обладать меньшими техническими возможностями достаточными лишь для поддержания собственных библиографических и полнотекстовых баз данных. Необходимым требованием для АИУ является прямой канал связи с одним из региональных ОИУ желательно не хуже выделенной телефонной линии. Опорные и ассоциированные узлы образуют единое информационное пространство, к которому получают доступ различные категории пользователей.

В плане доступа к информационным ресурсам базовых центров выделяются следующие уровни:

Уровень кампуса, или внутренней сети базового центра – в этом случае используются скоростные надежные коммуникации (FDDI, FastEthernet, SwitchedEthernet).

Уровень города и региона (межбиблиотечное взаимодействие). В настоящее время можно констатировать, что только в Москве и Санкт-Петербурге существует приемлемая территориальная сеть на базе ВОЛС, обеспечивающая надежное и быстрое соединение. В этом случае предполагается использование средств, предназначенных для уровня кампуса. Остальные города России, за редким исключением, характеризуются отсутствием надежных коммуникаций, что определяет особые требования для ПО ассоциированных библиотечных центров (например, они должны вести свои локальные БД, буферировать и защищать информацию). В 2004-2005 годах в Оренбургском научном центре УрО РАН были проведены работы по прокладке и подключению оптоволоконного кабеля и необходимого оборудования. На текущий момент для научных сотрудников филиалов и институтов центра наложен надежный и быстрый доступ к ресурсам Интернет.

Глобальный уровень. Предполагает взаимодействие с центрами, обладающими IP соединениями как в России, так и зарубежом. Планируется использование поисковых систем (Search Engine) и стандартных средств доступа в Internet. Этот уровень предполагает использование шлюзов к региональной библиографической БД. По мере развития Internet в регионе этот механизм доступа в открытые библиотечные сети будет становиться доминирующим.

Проектом предусматривалась разработка базового программного обеспечения, которое включает в себя проектирование информационных баз данных различных уровней и средств доступа к библиографической и полнотекстовой информации для различных категорий пользователей. Принятые архитектурные и технические решения должны были отвечать открытым стандартам и не накладывать ограничений на тип используемых платформ, обеспечивая при этом максимальную открытость системы. Технологический базис проекта: технология клиент/сервер, Internet-технологии, распределенные вычисления, открытые стандарты.

Для создаваемых информационно-поисковых систем (ИПС) характерны: архитектура – клиент/сервер, ядром которой является реляционная БД; многопользовательский одновременный online – доступ к БД; высокий уровень сложности схемы БД; большой объем БД; высокие требования к обеспечению целостности данных; высокая эффективность реализации поисковых запросов; высокий уровень мобильности и открытости. Высокий уровень сложности схемы БД ОНЦ УрО РАН требует использования централизованного управления метаданными. Для этого применяется активный словарь данных, который поддерживает обработку метаданных, как различных языков программирования, так и метаданных СУБД.

На этапе концептуального дизайна в качестве концептуальной модели данных используется ER модель (модель сущность-связь). Выбор этой модели объясняется тем, что она обладает достаточной общностью для отображения структуры и ограничений, свойственных реальной предметной области, и легко отображается в реляционную модель данных логического уровня.

На этапе логического дизайна осуществляется нормализация схемы БД, осуществляется транзакционный анализ. Проведение нормализации в сочетании с применением табличных

ограничений и ограниченный преобразования данных обеспечивает высокий уровень целостности данных ОНЦ УрО РАН. Ограничения преобразования данных реализованы с помощью механизма триггеров, которые обеспечивают целостность при вводе, удалении и модификации данных, а также выполнение контрольно-блокирующих функций и специальных функций (таких как авторизация, ведение истории изменений, архивация).

Большой объем данных при высоком уровне сложности схемы БД требует проведения на этапе физического дизайна обширной экспертизы схемы хранения БД, которая предполагает сложную взаимосвязанную настройку способа хранения кортежей, таблиц, индексов в зависимости от объема таблиц, критических транзакций, количества пользователей и т. п. Для обеспечения высокой эффективности выполнения поисковых запросов, в которых происходит соединение 2-х и более отношений, каждое из которых также является декартовым произведением 2-х и более отношений, выполнение последней операции соединения перенесено с сервера на клиентские машины. Это позволило увеличить скорость отработки пользовательских запросов в 6-7 раз.

В программное обеспечение ИПС входят:

- интерфейс пользователя на клиентской машине (IBM PC);
- динамическая библиотека, обеспечивающая логику взаимодействия с SQL-сервисом посредством использования API-функций и БД на сервере.

Интерфейс пользователя унифицирован и реализуется доступными средствами клиентских систем.

Краткая характеристика работы:

1. 01.04.2005-01.07.2005 – Разработка интерфейса и написание программных модулей поисковой системы.

2. 01.07.2005-01.10.2005 – Разработка конверторов и программных модулей взаимосвязи с другими библиотечными системами.

3. 01.10.2005-01.12.2005 – Интегрирование системы в пользовательские программы. Первичное наполнение системы информацией.

Структура системы:

1) Систему условно можно разделить на 2 структурные части: База Данных, набор скриптовых программ для работы с Базой Данных. Эти две части не могут функционировать независимо, но обеспечивают реализацию работы систе-

мы на различных ее этапах. Так набор скриптовых программ отражает этап обработки информации, ее ввода и вывода. База данных является только хранилищем информации. Такое разделение программного средства на Базу Данных и набор обслуживающих программ характерно для сетевых систем с большим объемом хранимой информации. Это объясняется тем, что для эффективной работы с большими объемами данных требуется сложная и большая по ресурсо-емкости подсистема хранения данных. Хорошим решением в таком случае является использование широко распространенных Систем Управления Базами Данных (СУБД). Так как в задачи системы входит предоставление доступа к данным широкому кругу пользователей, то основной средой работы системы выбрана глобальная сеть Internet. Этим выбором обусловлен и выбор средств программирования. Так основным языком программирования является скриптовый язык PHP4. Он поддерживается большинством Web-серверов и удобен для программирования Web-сервисов. В качестве СУБД выбрана MySQL4.1. Эта СУБД также поддерживается большинством Web-серверов и является одной из самых выгодных по соотношению эффективность работы/требование к ресурсам. Также язык программирования PHP4 имеет встроенные модули для взаимодействия с СУБД MySQL.

2) Систему можно разбить на две функционально не зависимые части: ввод и корректировка данных, поиск информации и ее отображение. Эти части зависят друг от друга только по данным, т. е. подсистема поиска работает с данными занесенными в БД подсистемой ввода. Обе подсистемы реализованы в виде скриптовых программ. Причиной такого разделения является изоляция друг от друга двух подсистем работы с данными: только чтение, чтение и запись. Это сделано для повышения безопасности работы с данными и предотвращение возможных попыток взлома системы. Если подсистема поиска и просмотра данных доступна всем пользователям Internet, то система корректировки и добавления данных доступна только пользователям сети УрО РАН.

Структура Базы Данных:

Так как СУБД MySQL является реляционной базой данных, то структура данных представлена таблицами. База данных системы состоит из 6 связанных таблиц. Ниже приведена схема базы данных отражающая взаимосвязи между таблицами (см. рис. 1).

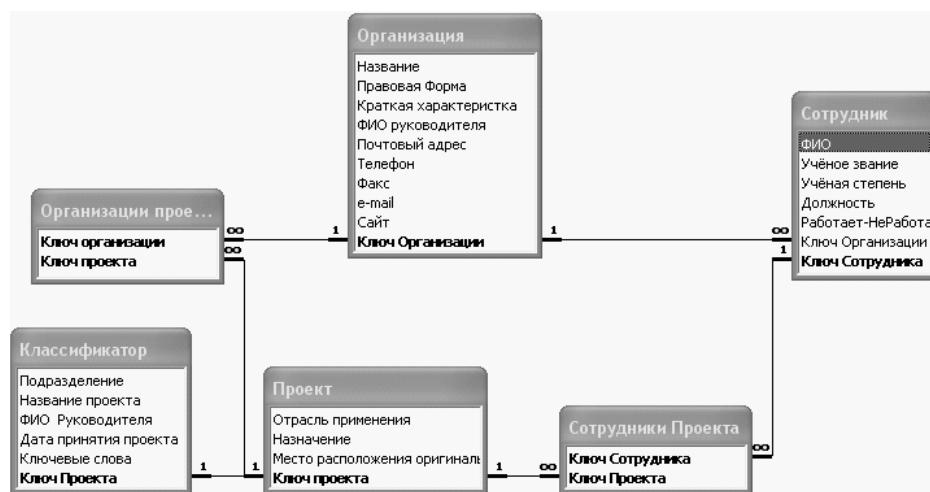


Рисунок 1. Схема взаимосвязи баз данных разработанного проекта

Структура программы:

Программа состоит из следующих модулей:

- **action.php** – этот модуль является основным модулем, который вызывается пользователем. Модуль в свою очередь подключает дополнительные части программы из других файлов. Модуль реализует интерфейс редактирования и добавления проектов.
- **search.php** – главный модуль поисковой системы. Объединяет дочерние модули.
- **db_modul.php** – модуль базы данных. Содержит типовые запросы на диалекте СУБД MySQL 4.1 к разработанной базе данных.
- **define.php** – модуль содержащий объявления глобальные переменных и констант.
- **functions.php** – этот модуль содержит часто используемые участки кода. Они вынесены в отдельный модуль для лучшей читаемости программы.
- **default_page.php** – содержит страницу, формируемую по умолчанию.

Также к программе можно отнести набор шаблонов страниц. Они содержат структуру основных страниц на языке разметки HTML 4 и

мета данные (метки вставки и изменения структуры страницы). Эти шаблоны используются как скелет для построения динамической страницы отсылаемой пользователю.

Так как изначально ОНЦ УрО РАН не имел собственного сайта, базой для размещения поисковой системы стал сайт Оренбургского филиала ГУ Институт экономики УрО РАН www.ofinos.ru (структурного подразделения ОНЦ УрО РАН). Сайт был создан в 2004 году и имел дизайн и адаптацию для целей филиала. Но для целей разработанного проекта сайт пришлось кардинально изменить и, в августе 2005 года, он был полностью переработан.

Таким образом, представленная информационная система созданная с целью автоматизации и централизации учета проектов (патентов, инноваций и т. д.) единой для всех подразделений ОНЦ УрО РАН. Это позволяет единым образом организовать доступ к информации о проектах. Так же позволило обеспечить предоставление системой удобного доступа к информации о проектах широкого круга пользователей.