Новые механизмы поиска на разнородных массивах данных. Из опыта внедрения библиотечных порталов

Достовалов Сергей Сергеевич, ведущий программист, Центр информационно-библиотечных систем Санкт-Петербургского политехнического университета Петра Великого

В докладе рассмотрены проблемы, технологии и механизмы поиска и индексирования разнородных массивов данных. Предложены решения для построения иерархии источников данных, представлены интерфейсы отображения и агрегации результатов поисковых запросов как из внутренних, так и внешних провайдеров на примере фундаментальной библиотеки СПбПУ.

Введение

Глобализация фонда – неизбежный процесс современной библиотеки. Этому процессу способствует интеграция библиотеки в консорциумы и объединения, подписки на БД электронных ресурсов, интеграция в Discovery сервисы. В результате такой интеграции библиотека становиться «обладателем» огромного количества ресурсов. Возникает проблема, как донести информацию об этих ресурсах пользователю? Как сделать так, чтобы пользователи не только знали о них, но и пользовались ими? Проблема усугубляется еще и тем, что сами сотрудники не всегда знают, на что они подписаны, вследствие чего не могут проинформировать пользователей.

Классические решения, к которым прибегают библиотеки для раскрытия своего фонда:

- размещения на портале поисковой формы, в которой перечислены базы и электронные каталоги;
- создание разделов, страниц, таблиц, объявлений в которых перечисляются подписные ресурсы и БД.

Казалось бы, задача решена, портал предоставляет информацию о ресурсах и некоторые механизмы поиска. Но проблемы остаются:

- пользователи могут не увидеть объявлений/списков подписных ресурсах – неудачно структурированная или устаревшая информация, типичное нежелание/лень пользователя исследовать портал;
- пользователи не знают, как ими пользоваться – каждый ресурс предлагает свои интерфейсы и механизмы поиска;
- количество поисковых форм, каталогов, внешних сервисов настолько велико, что пользователи просто теряются во всём этом многообразии;
- классический распределенный поиск по базам не поддерживает сортировку результатов поиска, результаты выдаются по мере поступления из источников. Может оказать так, что полезная запись попадает в конец выдачи, лишь потому, что сервер вернул результаты позже всех.

Очевидно, что такое положение вещей не устраивают никого: пользователей, которые не могут найти нужную информацию, библиотеку, тратящую деньги на ресурсы, которыми никто не пользуются.

В докладе рассмотренные обозначенные проблемы с точки зрения поисковых технологий и предложены решения в виде новых поисковых механизмов и интерфейсов, которые облегчают процесс поиска релевантных результатов, а также позволяют проинформировать пользователя об источниках, в которых они были найдены.

Общая картина

Фундаментальная библиотека Санкт-Петербургского Политехнического университета Петра Великого в обеспечивает доступ к следующим ресурсам:

- электронный каталог: ~ 700 000 записей;
- электронная библиотека, в которую входят коллекции документов: ~ 20 000 записей;
- подписка на ЭПОС: ~ 50 000 записей;
- подписка на МАРС: ~ 2,5 млн. записей;
- базы ВИНИТИ: ~ 9 млн;
- ресурсы АРБИКОН: ~50 млн;
- подписка на электронные ресурсы и БД: 64 шт. Только лишь в EBSCO количество записей больше 400 миллионов.

Количество и разнообразие ресурсов впечатляет. Ясно одно – классические подходы поиска информации не эффективны для обработки такого количества ресурсов и источников.

Традиционная схема поиска ресурсов библиотеки

Классический портал библиотеки представляет собой набор поисковых интерфейсов, ссылок на внешние ресурсы и БД. Примерную схему взаимодействия ресурсов, которыми владеет ФБ СПбПУ и портала можно изобразить следующим образом:

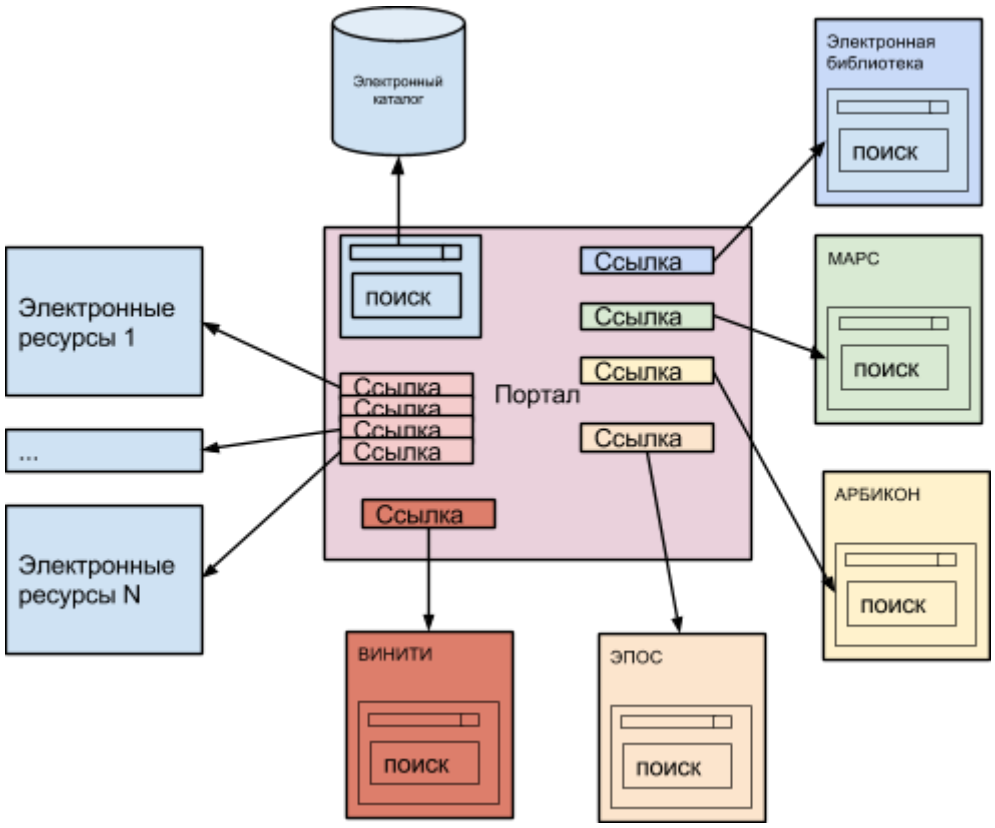

Рис. 1. Схема взаимодействия ресурсов и портала библиотеки

Из схемы, представленной выше, видно, что для того чтобы пользователь смог поискать по всем ресурсам, ему нужно перейти как минимум в 7 поисковых интерфейсов. В реальности, с учетом количества подписок на БД электронных ресурсов, этих интерфейсов может быть около 50. Сколько это может занять времени у одного человека можно только догадываться.

Сервис Discovery

Для минимизации схемы, изображенной на рис. 1, было бы логично объединить все источники под единым поисковым интерфейсом.

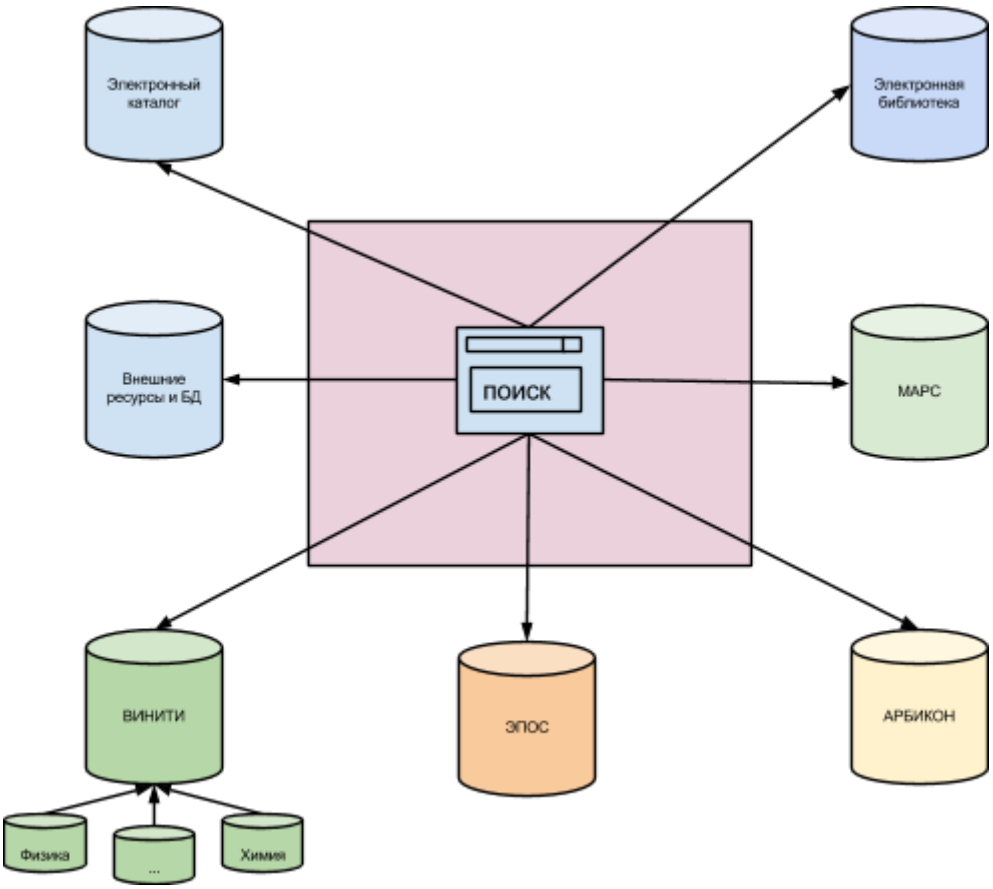

Рис. 2. Схема единого поиска по ресурсам

На рис. 2. изображена схема единого поиска. Такая схема была бы идеальна с точки зрения удобства пользования, но она порождает следующие проблемы:

- **способ взаимодействия**: не все источники поддерживают стандартные протоколы и форматы;
- **скорость поиска**: каждый источник обладает разным количеством записей, и обслуживается на разных вычислительных мощностях. Время реакции может исчисляться от долей секунды до нескольких минут;
- **ранжирование и фильтрация результатов**: некоторые источники не поддерживают ранжирование, другие поддерживают, но способы вычисления коэффициента релевантности (КР) различаются, к тому КР зависит от алгоритма вычисления, количества документов и проиндексированных атрибутов. Каким образом объединить поисковую выдачу из разных источников, чтобы наиболее релевантные результаты попали в начало выдачи?
- **дублетные результаты**: ликвидация дублетов в реальном времени практически невозможна на больших объемах результирующей выборки.

Всё это делает схему распределенного поиска мало пригодной для комфортной работы и получения адекватного результата.

Сервис по типу Discovery предполагает решение вышеописанных проблем. Основные этапы работы сервиса Discovery:

- предварительная загрузка записей из разных источников;
- индексирование записей в едином индексе;
- предоставление пользователю поисковый механизм, в котором они могли бы искать записи используя общие алгоритмы ранжирования и фильтрации.

Но Discovery сервисы так же не лишены проблем:

- для сбора записей необходимо поддержать различные форматы и механизмы и протоколы сбора записей;
- не все записи являются корректными. Необходимо производить проверку записей, при этом учитываю формат и схему и особенности их составления;
- хранение большого количества записей предъявляет требования к вычислительной мощности сервера, вплоть до принятие мер по распределению хранилища по разным серверам (параллельное масштабирование);
- поиск и индексирование большого количества записей так же может потребовать распараллеливания индекса серверам;
- большое количество записей источника (сотни миллионов), авторские права, лицензирование доступа, делают сбор записей невозможным в принципе, вследствие чего источник не будет участвует в поиске.

Кроме технических проблем, есть еще и проблемы, которые затрагивают идентичность библиотеки:

- при увеличении количества записей, записи библиотеки теряются в огромном количестве записей других библиотек;
- отражение структуры каталогов и коллекций библиотеки. В качестве примера можно привести коллекции в электронной библиотеке;

В случае, если библиотека пользуется внешними Discovery сервисами, пользователи вынуждены «уходить» с портала, что делает проблематичным или невозможным предоставления локальных услуг и сервисов, например, забронировать документ, сохранить в избранном и т.п.

В результате анализа потребностей ФБ СПбПУ, и учитывая вышеизложенные проблемы, были сформулированы требования к поиску:

- предоставить единое окно поиска, объединив различные источники и базы данных, при этом, учитывая специфику структуры каждого источника;
- в случае, если объединение источника невозможно вследствие его технической сложности, большого объема, лицензионной политики, предоставить механизм параллельного поиска, который бы учитывал структуру и специфику источника, при этом был бы удобен с точки зрения удобства использования и навигации, а так же обеспечивать приемлемое время обработки запроса пользователя.

Сервис Руслан-Discovery

Для реализации выдвинутых требования был применен продукт Руслан-Discovery. Архитектура сервиса представлена на рис. 3.

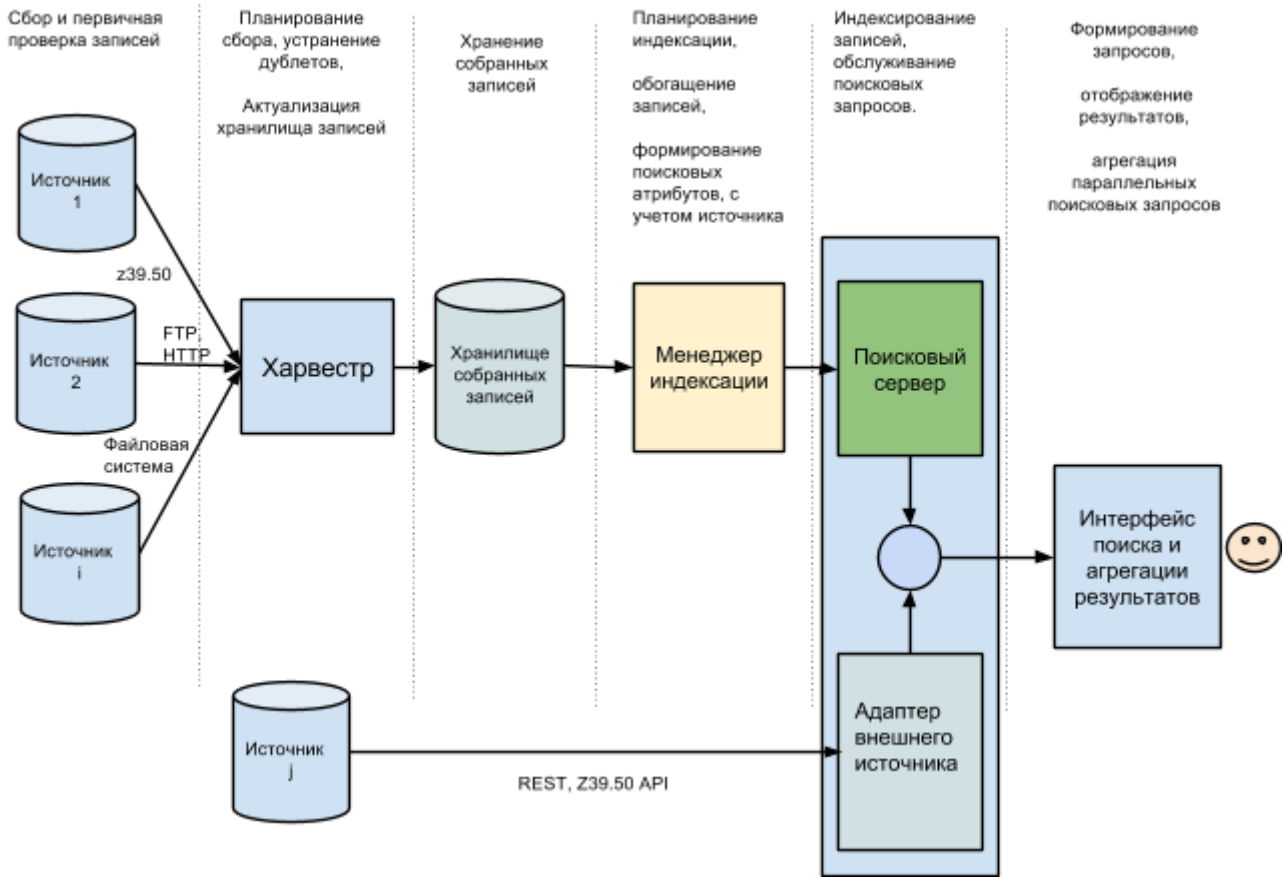

Рис. 3. Архитектура сервиса Руслан-Discovery

Основные этапы работы сервиса:

1. Сбор записей из внешних источников, используя стандартные протоколы и форматы. Поддержка различных политик сбора;
2. Хранение записей в хранилище, с возможностью устранения дублетов. Актуализация записей по расписанию или по запросу.
3. Индексирование записей с учетом спецификации источника.
4. Обслуживание поисковых запросов от пользователей.
5. Интерфейс поиска направляет запросы к различным источникам и агрегирует результаты поиска.

Результаты работы

После внедрения сервиса, схема взаимодействия ресурсов и портала библиотеки приняла следующий вид:

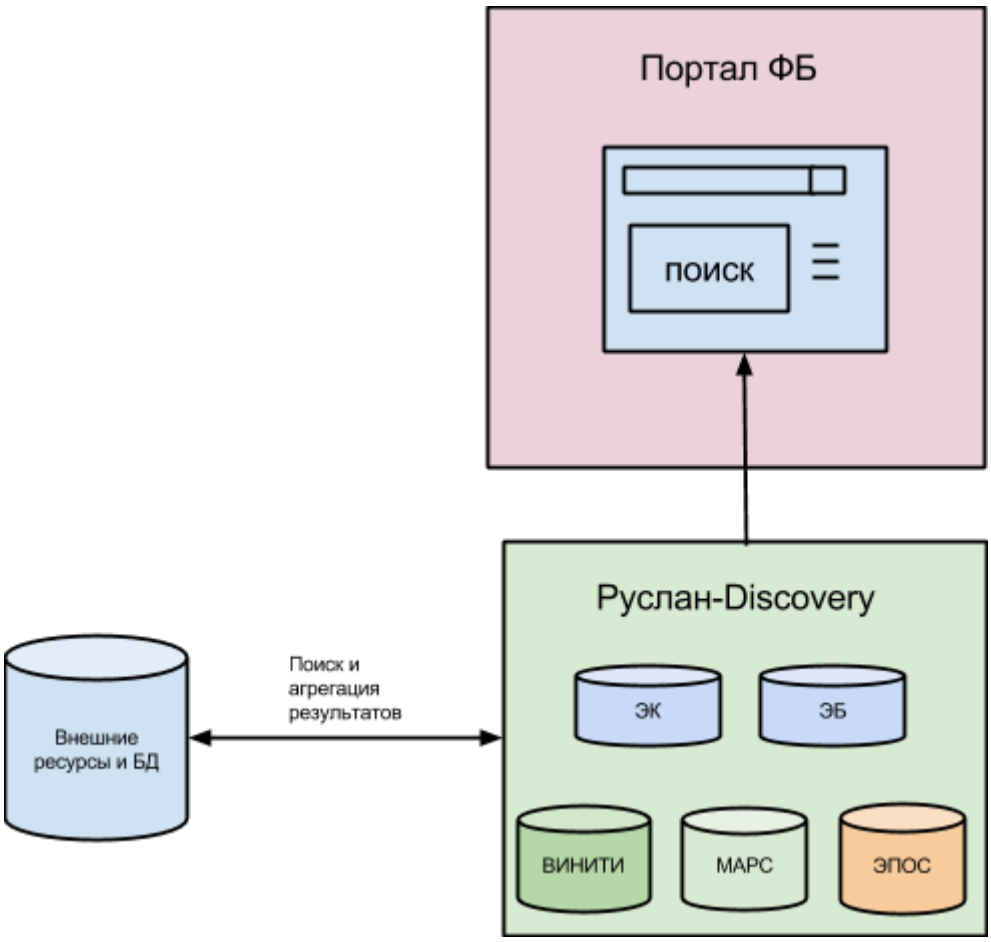

Рис. 4. Схема единого поиска с использованием discovery сервиса

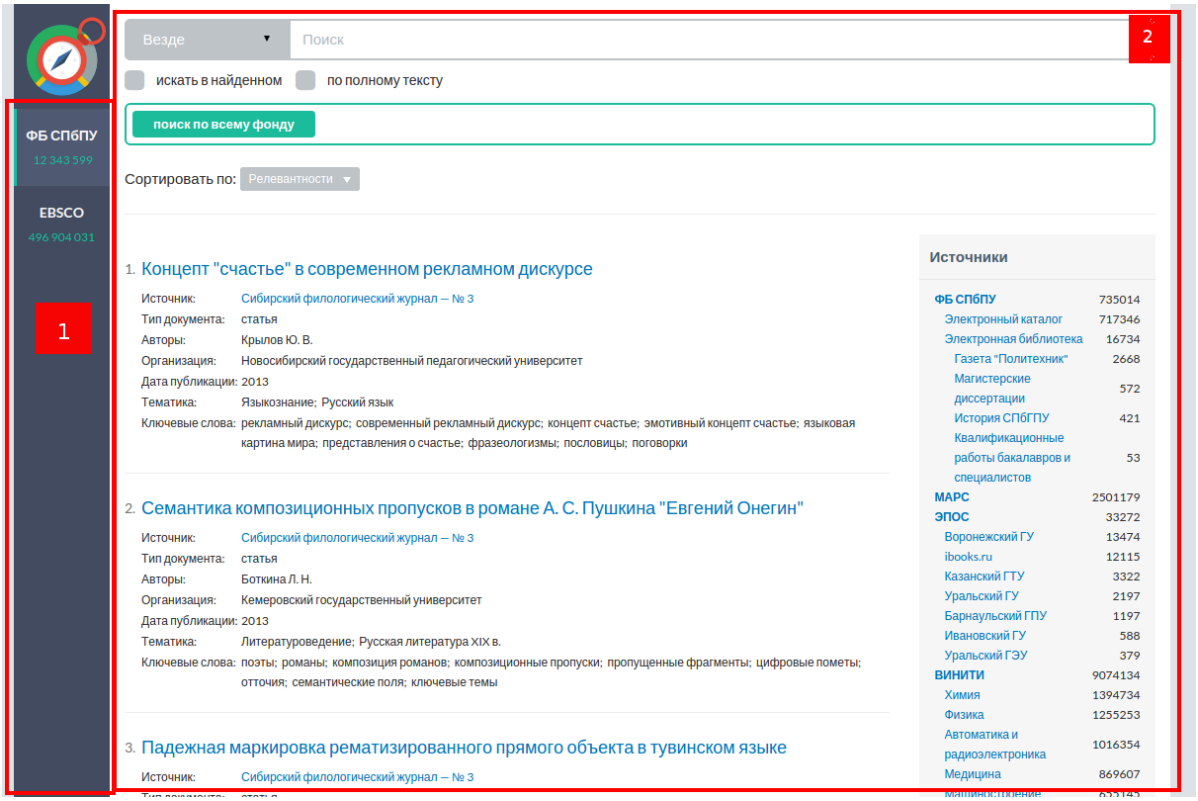

Рис. 5. Единый интерфейс поиска

На рисунке 5 представлен интерфейс единого поиска, который является агрегатором результатов различных сервис-провайдеров. Контурами отмечены области:

1. информационная панель, в которой перечислены поставщики Discovery сервисов с информацией о количестве найденных результатов;
2. интерфейс текущего discovery сервиса. В данном случае внутренний сервис, обслуживающий ФБ СПбПУ.

Основное преимущество интерфейса в том, что пользователь сразу видит общую картину поиск. А так же, что немаловажно, информирует пользователя о том, на какие ресурсы подписана библиотека, попутно осуществляя поиск в этих ресурсах.

Одним из основных требований к интерфейсу является удобство навигации по внешним источникам не покидая портала. Для удовлетворения этого требования интерфейс поиска был реализован в виде RIA (Rich Internet Application), что позволило переключаться между источниками не теряя поискового контекста, при этом поддерживая параллельный особенности поиска и услуг каждого сервис-провайдера.

Поддержка особенностей сервис-провайдера осуществляется благодаря системе модулей, каждый из которых поддерживает особенности фильтрации, отображения и протокол взаимодействия (API). На рисунках 6 и 7 отображена работа модулей, обслуживающих сервис провайдеры ФБ СПбПУ и EBSCO:

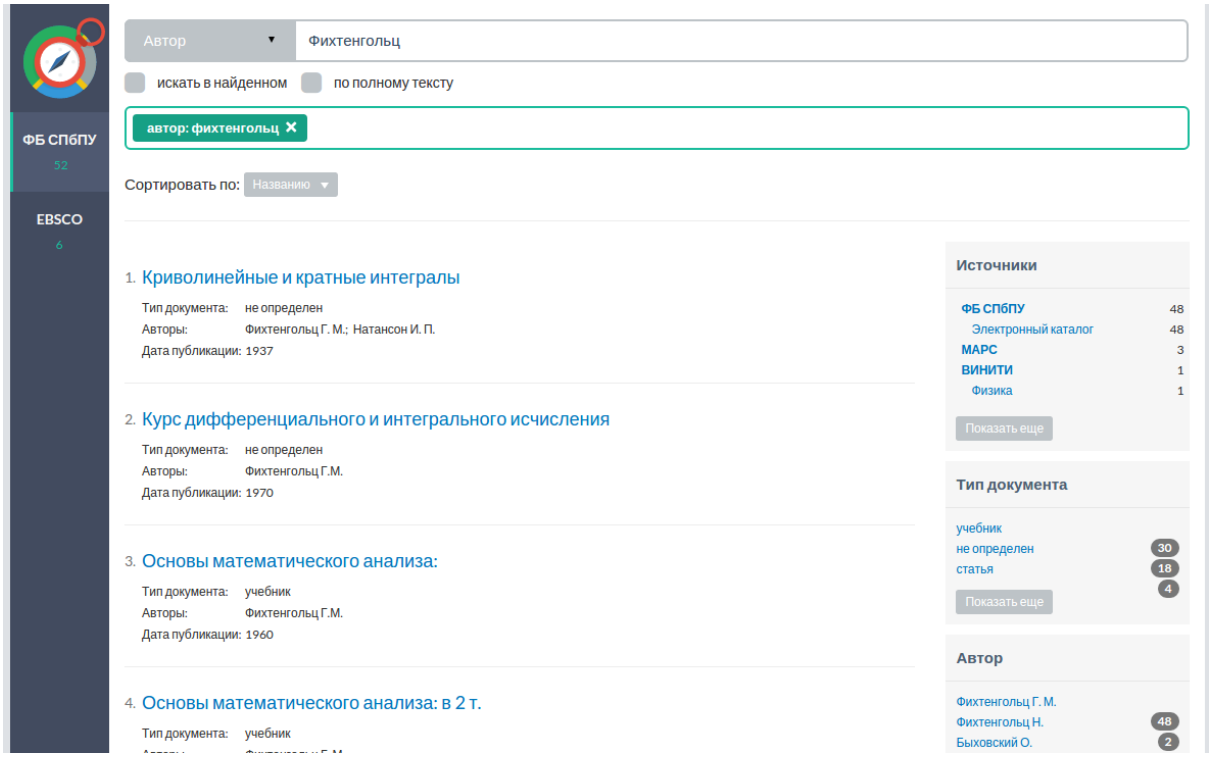

Рис. 6. интерфейс сервис-провайдера ФБ СПбПУ (Руслан-Discovery)

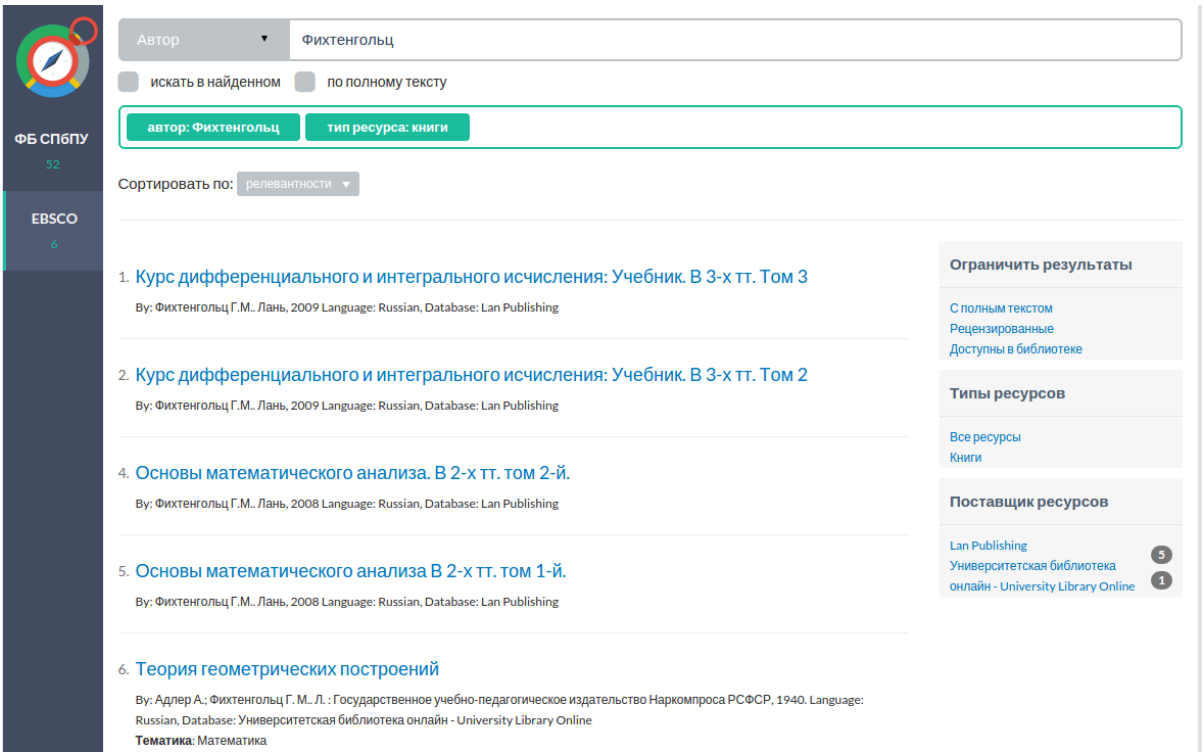

Рис. 7. интерфейс провайдера EBSCO

Из рисунков 6 и 7 видно, как поисковый интерфейс и фильтры адаптируются под конкретного провайдера.

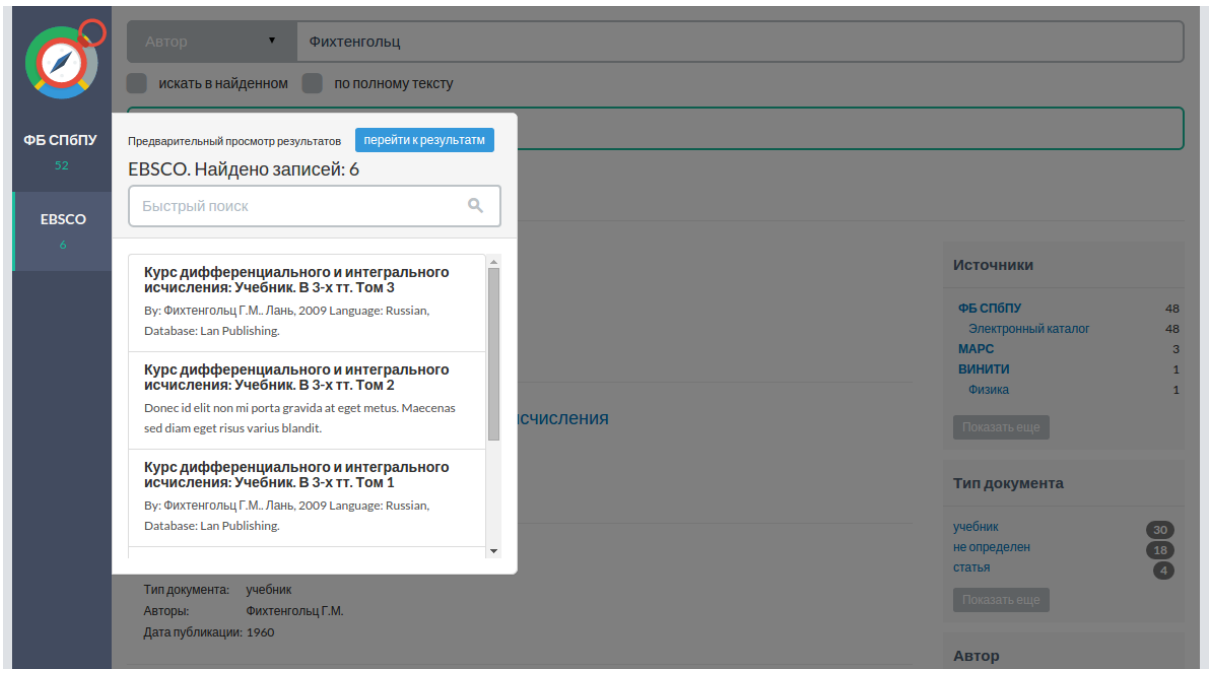

Рис. 8. Предварительный просмотр результатов

Еще одной концепцией единого поиска является предварительный просмотр результатов сервис-провайдера (рис. 8), что позволяет быстро оценить результаты поиска не переключаться на экран провайдера.

Немаловажным аспектом является поддержка локальных сервисов библиотек. На рис. 9 показан экран детальной информации о записи, которая была найден в электронном каталоге ФБ СПбПУ. Помимо общей информации о записи, отображено местоположение экземпляров, с сопутствующими услугами бронирования, заказа по МБА, сохранения в «Избранном».

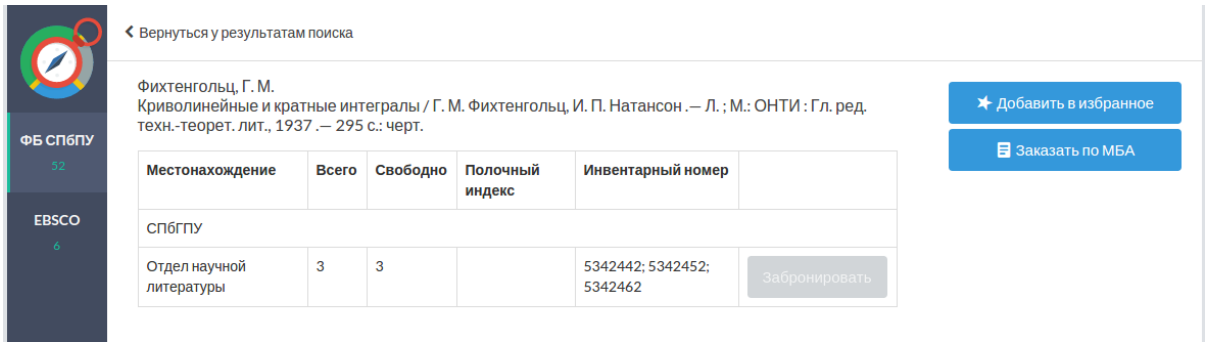

Рис. 9. Экран детальной информации о записи

Заключение

Попытки объединения и представления разнородных источников информации осуществлялись давно. Основными препятствиями к реализации удобной навигации и поиска были вычислительные возможности серверов и возможность создавать динамические приложения для браузеров. Проблема усугублялась еще и тем, что каждый браузер предлагал своё видение технологий, а некоторые вообще этих технологий не поддерживали, в качестве примера, кон-

серватизм Internet Explorer’a. Бум облачных сервисов и приложений заставил производителей браузеров и инструментов веб-разработки ускорить процесс внедрения новых технологий, что позволило в некоторой мере пересмотреть стандартные подходы к построению поискового интерфейса, не теряя уверенности, что он перестанет работать в каком-либо браузере.

Развитие поисковых возможностей Руслан-Discovery позволило объединить источники в иерархические группы. Это позволило выделить не только отдельные базы, но и коллекции, а также подколлекции, входящие в эти базы.

Список литературы:

1. Достовалов С. С., новая "одежда" для библиотечного каталога - система поиска нового поколения [Электронный ресурс] // Корпоративные библиотечные системы: технологии и инновации: X Междунар. науч.-практ. конф. и выст., http://elib.spbstu.ru/dl/2577.pdf.
2. Литвинова Н.Н. Проблемы внедрения поисковых сервисов типа Discovery в библиотеках // Библиотековедение. – 2013. – №6. – с.41-45.
3. Sadeh, T. From search to discovery// proceedings of IFLA WLIC 2013, Singapore, Режим доступа:: http://library.ifla.org/104/1/098-sadeh-en.pdf.