2002 | Том 5 | Выпуск 1

О реализации электронной библиотеки с использованием протоколов HTTP и Z39.50

Максимов Н.В.
РГГУ

Сысойкина М.А.
ИНИОН РАН

Статья посвящена вопросам реализации информационной системы, предоставляющей унифицированный доступ к распределенным ресурсам. В данной работе рассмотрены общие вопросы организации доступа к распределенным ИР, а также основные компоненты такой системы. Особое внимание уделено использованию различных типов информационных ресурсов: локальных, удаленных, ассоциированных. Раскрываются основные аспекты совместного использования протоколов Z39.50 и HTTP для доступа к распределенным ИР. В качестве примера приводится опыт создания подобной системы на базе информационного пространства научной информации ИНИОН РАН с использованием ИПС IRBIS.

Введение

Интенсивное развитие и широкое использования электронных библиотек (ЭБ) уже сегодня привели к тому, что огромное количество разнородной информации стало доступно пользователям. Но по мере создания все новых и новых ЭБ и онлайновых каталогов, сформировался целый ряд требований к такого рода информационным ресурсам и способам их организации. ЭБ, по сути своей, предназначены для хранения разнородной электронной информации и должны предоставлять пользователям удаленный доступ к ней по сети. Но при наличии в настоящее время значительного числа стандартов представления и передачи электронных документов, непременным требованием к ЭБ становится возможность представлять информацию в унифицированной форме, что обеспечит более эффективное использование распределенных информационных ресурсов. Также большое значение для внедрения ЭБ имеет дружественный пользовательский интерфейс, предлагающий достаточно широкие возможности по поиску информации.

Интерфейс пользователя является одним из ключевых факторов при создании ЭБ. Реализация интерфейса во многом зависит от используемого в ЭБ протокола передачи данных. Например, если используется HTTP, то традиционно, клиентское приложение имеет Web-интерфейс, а если используется Z39.50, то возможно, клиентская программа будет реализована в виде обыкновенного графического интерфейса.

Следующим, также немаловажным, фактором являются средства, обеспечивающие представление информации, и возможность поддержки нескольких форматов выдачи документа пользователю. В случае использования протокола Z39.50 этот вопрос решается достаточно просто, так как выдача в различных форматах заложена в самом стандарте. Таким образом, клиент может получить один и тот же документ в нескольких видах, например, в виде простого неструктурированного текста, в ряде MARC-подобных форматов (на базе ISO 2709), или же в том виде, в котором документ хранится в БД на сервере.

И, пожалуй, последним фактором является организация процесса поиска, реализуемого в ЭБ. Сам процесс поиска во многом связан как с используемыми протоколами доступа, так и с реализацией пользовательских интерфейсов [1]. От используемого протокола зависит возможность или невозможность реализации таких функций, как поддержка поисковой сессии, итеративный поиск, разнообразие поисковых интерфейсов. В свою очередь набор реализуемых функций во многом влияет на сложность и эффективность реализаций поисковых стратегий.

Таким образом, можно видеть, что все вопросы, возникающие при реализации ЭБ, во многом взаимосвязаны между собой, и их решение в значительной степени определяется протоколами доступа.

Управление доступом к информационным ресурсам: протоколы HTTP и Z39.50

Управление доступом к распределенным информационным ресурсам и взаимодействие ЭБ осуществляется по принципу распределенных систем на базе открытых стандартов обмена данными. Для реализации ЭБ используются в основном два протокола: Z39.50 и HTTP. Протокол Z39.50, изначально ориентированный на информационно-поисковые задачи, функционально обеспечивает поддержку поисковой сессии.

Сессией является последовательность функционально связанных поисковых сеансов (операций), направленная на получение логически целостного результата, каждый из которых выполняется в рамках отдельной сетевой транзакции.

Причем поддержка многошагового поискового процесса осуществляется средствами Z-сервера, управляющего взаимодействием как на сеансовом, так и на сессионном уровне. Это значит, что каждый последующий запрос от клиента будет основываться на предыдущем ответе сервера. В ходе сессии, на сервере хранится информация обо всех выполненных запросах и их результатах.

Z39.50 - протокол прикладного уровня в рамках семиуровневой модели OSI, и поэтому может работать на различных типах сетей, независимо от транспортного уровня, а взаимодействие клиентского и серверного приложений осуществляется по модели "запрос-ответ".

Но, именно специализированность, требующая унификации представления ресурсов и наличия интерфейсных средств на стороне пользователя (клиента),

предопределила более широкое использование функционально более простого протокола HTTP, управляющего взаимодействием только на уровне отдельных транзакций (сеансов). В случае HTTP-протокола поддержка поисковых сессий может осуществляться специализированными средствами управления доступом электронной библиотеки на стороне сервера, или клиентской программой работы с ресурсом.

Протокол HTTP считается достаточно распространенным и удобным средством извлечения и представления информации. Однако сам HTTP не предоставляет возможностей по поиску документов в базах данных. Эта услуга может быть реализована только при использовании расширений протокола, например, CGI. При этом HTTP-сообщения представляют собой обычные текстовые документы, передаваемые по сети.

Что касается поддержки сессии, то отсутствие этого механизма связано не со спецификацией протокола вообще, а лишь с существующими его реализациями. Спецификация HTTP не оговаривает параметры соединения, то есть HTTP может выполняться над любыми пакетами протоколов, и, следовательно, может осуществлять более одной транзакции в рамках одного соединения. Серверу ничего не известно о результатах предыдущего взаимодействия с клиентом. Каждый раз, когда клиент собирается послать серверу запрос, он должен открыть соединение, которое закрывается сервером после посылки ответа.

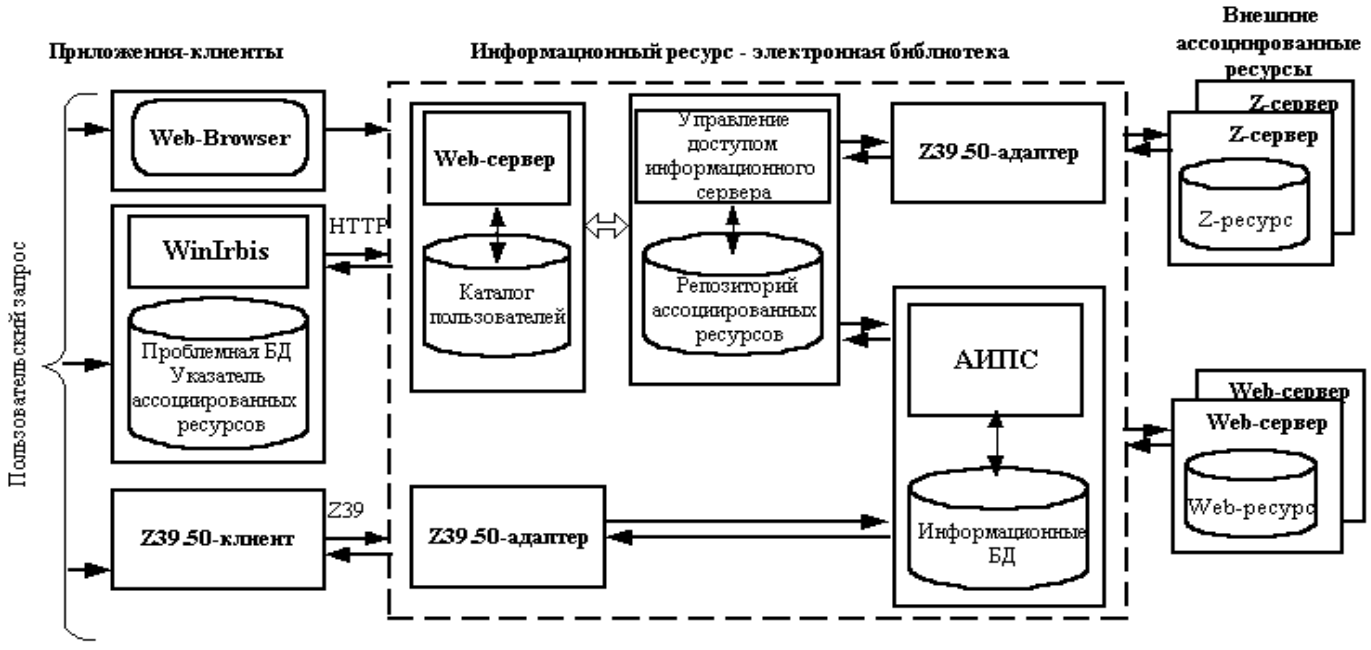

Рис 1.Общая схема доступа к ИР

Общая схема организации доступа к распределенным ИР представлена на рис. 1. Здесь предполагается использование ресурсов с различными вариантами размещения - локальные БД на компьютере пользователя; удаленные базы под управлением ЭБ; удаленные ресурсы, внешние по отношению к ЭБ. Доступ

клиента к ресурсам ЭБ возможен как по протоколу HTTP, так и по протоколу Z39.50, причем оба могут использоваться в одной сессии, например, для получения от Z-ресурса полного текста документа, найденного в ЭБ HTTP-средствами, или при поиске по запросу в ассоциированных ресурсах (тематически совместимых и, возможно, дополняющих основной ресурс, к которому явно обратился пользователь).

Компоненты ЭБ, управляющие доступом к ресурсам

В составе электронной библиотеки можно выделить три управляющих компонента:

- сервер связи, обеспечивающий доступ к программно-информационному комплексу ЭБ;
- собственно информационный сервер - информационно-поисковая система, обеспечивающая поиск и выдачу информационных объектов ЭБ;
- сервер управления доступом пользователей библиотеки к информационным ресурсам ЭБ.

Сервер связи обеспечивает сетевое взаимодействие клиентской программы с программным обеспечением информационного сервера.

Информационный сервер - это хранилище данных, представляющих коллекции разнородных логически или физически взаимосвязанных документов (представляющих компоненты трехуровневой модели ЭБ), доступ к которым обеспечивается средствами информационно-поисковой системы. При этом поисковые (реферативные или библиографические) и метаинформационные компоненты реализуются в виде структурированных баз данных, в то время как компоненты уровня первичной информации могут быть представлены как полнотекстовыми БД, так и коллекциями файлов, каждый из которых обычно хранит в том или ином графическом или печать-ориентированном формате образ страниц документа.

Сервер управления доступом к информационным ресурсам, реализующий ограничения доступа к компонентам электронной библиотеки, является специализированной системой, обеспечивающей поддержку сессионного уровня поискового процесса, а также возможности административного учета и управления доступом на уровне отдельного пользователя к отдельному информационному и функциональному ресурсу. Реализуется в виде специализированных баз данных, в том числе:

- каталога информационных ресурсов, хранящего сведения о локализованных и ассоциированных ресурсах;
- каталога пользователей, хранящего учетные данные о каждом пользователе и список доступных ему информационных ресурсов;
- технологической базы данных сеансов, содержащей информацию о

текущем состоянии сеанса каждой поисковой сессии;
- базы данных учетно-статистической информации об использовании ресурсов.

Репозиторий (указатель) ассоциированных ИР формируется путем выделения проблемно-ориентированных ресурсов на основе автоматизированного анализа информационной совместимости (определением меры лексической совместимости профилей пользователя и ресурса).

Такие средства позволяют определять пространство доступных ресурсов для каждого отдельного пользователя, защищая при этом определенные базы данных и/или массивы библиотеки от несанкционированного обращения.

Использование ассоциированных ресурсов осуществляется либо путем перехода в среду самого ресурса, либо обращением к отдельным его объектам через ссылки, размещенные в операционном (интерфейсном) пространстве текущего ресурса.

Доступность информационных объектов можно рассматривать с двух точек зрения: с точки зрения пользователя и с точки зрения администратора. Первая точка зрения - это взгляд "снаружи", отражает особенности объектов, указываемых ссылками, и, соответственно, определяет способы их использования. С другой стороны (с точки зрения администратора), объекты и связи можно разделить по способу управления на две группы - те, которые находятся под управлением файловой системы, и - под управлением СУБД. В первом случае обращение производится по адресу-идентификатору, а во втором - по совокупности значений атрибутов. В свою очередь объект может быть локальным (в той же вычислительной или операционной системе) или удаленным. Кроме того, такие связи (указатели ресурсов) должны быть максимально независимы от изменений, например, генерироваться (в виде указателей-идентификаторов объектов или условий отбора по атрибутам) динамически в момент визуализации документов.

Можно выделить [2] несколько типов связей между информационными объектами:

- "гипертекстовые" связи, указывающие на записи той же базы данных, например, на документы того же автора;
- "гипертекстовые" связи, указывающие на записи другого информационного уровня, например, на расшифровку кода рубрики - в базу данных рубрикатора, или наоборот;
- связи-запросы, позволяющие инициировать поиск ассоциированных документов по его атрибутам во внешней (удаленной) базе данных, например, поиск на удаленном сервере полных текстов по библиографии, найденной в локальной БД;
- связи-указатели на локальные документальные ресурсы (документы-аналоги или, например, файлы полных текстов документов, размещенных в хранилище той же вычислительной среды).

Совместное использование потоколов Z39.50 и HTTP для доступа к ассоциированным ресурсам

Перечисленные методы и средства обеспечивают систематизированность информационного пространства и возможность разнопланового вхождения в предметную область (через рубрикаторы, тематический поиск и т.д.), а также эффективные технологии управления доступом к ассоциированным ресурсам, в том числе, например, переадресацией запроса для расширенного поиска, предоставлением электронной копии полного текста документа или ее поиском в соответствующем хранилище.

Например, информационное пространство научной информации по общественным наукам ИНИОН РАН включает базы данных вторичной реферативно-библиографической информации, массив электронных копий полных текстов статей, а также справочные базы тезаурусов и рубрикатора. Эти ресурсы предоставляются пользователю либо как электронная библиотека (под управлением WebIrbis), либо на компакт-дисках (под управлением WinIrbis) - как ежегодно обновляемые выпуски ретроспективных проблемно-ориентированных баз данных.

Информационная система IRBIS предназначена для многоаспектного поиска информации в базах данных разнородных документов. Система работает в среде MS DOS (Irbis), MS Windows (WinIrbis) в локальном или сетевом режиме, в том числе в Word Wide Web (WebIrbis) со шлюзом Z39.50. Система и технологии разработаны авторским коллективом в составе: Васиной Е.Н. (1991-2001), Голицыной О.Л. (1989-2001), Ивашко В.Г. (1989-1992), Максимова Н.В. (1989-2001), Петрова Ю.В. (1989-1991), Резниченко П.И. (1995-2001), Сысойкиной М.А.(2000-2001) при поддержке СКБ ВИНИТИ РАН, ИНИОН РАН, ВНТИЦентра РФ. Все права зарегистрированы (Свидетельства о регистрации РосАПО с 1992г).

Оба варианта доступа обеспечивают сессионный уровень поискового процесса: протокол поиска позволяет идентифицировать каждый информационный поиск в целом, обеспечивая при этом выборочное использование каждого промежуточного результата. При этом, пользователь, работающий с локальной базой в интерфейсной среде WinIrbis, имеет возможность продолжить сессию в ЭБ, например, для поиска в обновленной части базы, или получить полный текст. Аналогично, пользователь электронной библиотеки, обращающийся из стандартного Web-броузера, имеет возможность переадресовать запрос в ассоциированный ресурс, в том числе по протоколу Z39.50.
Достаточно большой интерес на сегодняшний день представляют собой вопросы взаимодействия в электронных библиотеках, а также совместного использования Z39.50 и HTTP.

В общем случае работу системы, предоставляющей возможности взаимодействия по протоколам Z39.50 и HTTP можно представить следующей схемой:

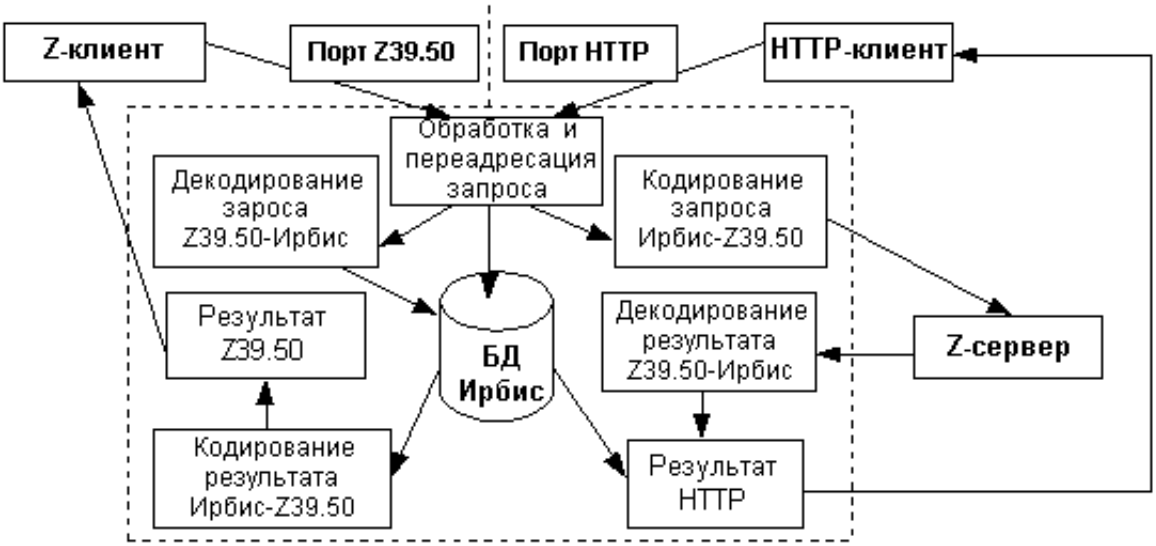

Рис.2 Система предоставления доступа к ресурсам по протоколам Z39.50 и HTTP.

Такая система предоставляет доступ к локальным и удаленным ресурсам как по протоколу HTTP так и по протоколу Z39.50. Причем доступ к удаленным Z-серверам может осуществляться как путем переадресации запроса, поступившего по HTTP, так и в результате поиска ассоциированных ресурсов.

В последнем случае, при поиске ассоциированного ресурса по протоколу Z39.50 реализуется механизм связей-запросов. То есть, при выдаче пользователю вторичного документа (например, библиографического описания статьи) генерируется ссылка для вызова модуля, передающего соответствующие параметры Z-клиенту, который в свою очередь осуществляет взаимодействие с указанными Z-серверами.

Для формирования такого запроса используется указатель ассоциированных ресурсов, (в данном случае список доступных Z-серверов и массивов информации на этих серверах), а также спецификации критериев отбора, используемых поисковыми системами этих серверов. Основным условием выбора сервера в качестве ассоциированного ресурса естественно является его тематическое соответствие. В качестве примера рассмотрим следующие:

- наличие нужного ресурса (например, полного текста документа) на удаленном сервере точно известно. Такая ситуация возможна, если БД на локальном сервере содержит описания, допустим, ряда сериальных изданий, а на удаленном сервере поддерживается массив полных текстов статей этих же изданий.
- наличие нужного ресурса может определяться сходством тематик БД. В этом случае положительный результат поиска не всегда гарантирован.

Очевидно, что в массе своей информационные системы используют различные способы хранения информации в базах данных. Причем, различаются не только форматы хранения информации, но также форматы ее передачи и представления, а также способы доступа к этой информации (в частности, формы запросов к БД).

Несомненно, протокол Z39.50 считается довольно мощным унифицирующим средством как для доступа к данным, так и для их передачи. Однако, даже не смотря на поддержку многими системами протокола Z39.50 [3], существующие различия в его применении достаточно сильно влияют на организацию и средства обращения к хранимой информации.

Прежде всего, каждый сервер поддерживает какой либо свой формат выдачи документов. Отчасти это преодолевается тем, что каждый сервер по умолчанию должен поддерживать выдачу в простом текстовом формате (sutrs - неструктурированный текст), и, кроме того, большинство серверов поддерживают выдачу в формате USMARC.

Сложнее решить проблему единообразия представления запроса для поиска информации на нескольких серверах одновременно.

Как известно, любая система, работающая по протоколу Z39.50 должна обязательно поддерживать набор атрибутов Bib-1, определяющий способ указания терминов в поисковом запросе. В состав Bib-1 входят шесть групп атрибутов:

- набор Use - непосредственно поля, или как их называют "точки доступа";
- набор Relation - атрибуты отношений (например, >, <, = и т.д.);
- набор Position - атрибуты, определяющие местонахождение поискового термина в поле (первое слово в поле, в подполе, любая позиция)
- набор Structure - атрибуты, определяющие структуру поискового термина (слово, фраза, ключ и т.д.)
- набор Truncation - атрибуты, определяющие возможность усечения термина;
- набор Completeness - атрибуты, определяющие полноту поля или подполя.

С одной стороны, такая детализация в определении свойств поисковых терминов предполагает достаточно большую гибкость в составлении запросов, однако с другой стороны создает препятствия для поиска по одному запросу, обращаемому одновременно к нескольким серверам. Связано это с тем, что в силу использования различных средств управления базами данных на различных серверах практически невозможно реализовать все предлагаемые стандартом наборы атрибутов. Часто случается так, что для поиска на одном сервере указание какого-либо атрибута является необходимым условием для успешного поиска, а на другом сервере этот атрибут, или даже вся группа атрибутов просто не реализована. В этом случае невозможно будет использовать единый запрос для обращения к базам данных.

Особенно актуальна эта проблема в случае поиска, когда неизвестно точное местонахождение ресурса, и поиск должен проводиться на нескольких Z-серверах.

Прежде всего, при формировании запроса возникает проблема выбора поисковых полей (точек доступа), включаемых в поисковое выражение. Наборы точек доступа, поддерживаемых разными Z-серверами, часто различаются. Несмотря на то, что такие поля, как Автор, Заглавие или Дата издания поддерживаются

практически во всех базах данных, часто этого бывает недостаточно. А такие поля, как ISSN, например, могут и не поддерживаться сервером. Все это влечет за собой непредсказуемость успешности поиска.

Далее, абстрактная структура записи Z39.50 не всегда одинаково отображается на физическую структуру документа БД.

Например, поле Автор (атрибут Use - 1003) в одной реализации может содержать только первого автора статьи (аналогично полю 100 а записи USMARC), а в другой сразу всех авторов.

Аналогичная ситуация может возникнуть с полем Заглавие. Например, если на одном сервере заглавие и параллельное заглавие или сведения, относящиеся к заглавию разнесены по разным полям, то в реализации другого сервера все эти поля объединены под одной меткой.

И, наконец, различия могут наблюдаться в реализации дополнительных атрибутов из набора Bib-1. Наиболее ярким примером может служит набор атрибутов Structure, определяющий структуру поискового термина. Примером различий в этом случае может служить реализация поля Subject (атрибут Use - 21), соответствующий полю Ключевые слова. К содержимому этого поля при поиске можно обращаться по-разному - как к фразе, или как к списку слов. Соответственно и атрибуты набора Structure должны быть указаны по-разному для каждого случая.

Исходя из приведенных положений, в рамках реализации поиска ассоциированных ресурсов по протоколу Z39.50 на сервере ИНИОН РАН была реализована справочная структура, определяющая доступные ресурсы и необходимые параметры. Справочник состоит из двух таблиц, в первой из которых содержатся указатели на ресурсы. Нужный ресурс идентифицируется шифром хранения документа в библиотеке ИНИОН РАН и соответствующим ему ISSN. По шифру хранения определяется ISSN сериального издания, в котором опубликована искомая статья. При выдаче статьи, для которой в справочнике содержится запись о наличии полного текста, формируется ссылка, вызывающая модуль Z-клиента с соответствующими параметрами.

Помимо шифров ISSN изданий в первом справочнике хранится информация о самом Z-сервере, и БД. После определения сервера происходит обращение к справочнику точек доступа, где для найденного сервера определяется шаблон запроса. В связи с тем, что, как уже было сказано выше, наборы атрибутов могут значительно для различных серверов, да и для различных БД в рамках одного сервера, то решено было определить такой набор точек доступа, поддержка которых будет наиболее вероятна для большинства Z-серверов. При отборе также учитывалась необходимость однозначно идентифицировать документ, и, естественно, соответствие выбираемых точек аналогичным полям в базах данных ИНИОН. В результате было выбрано 7 точек доступа:

8 - Identifier-ISSN;
31 - Date-publication;

1003 - Author-name;
4 - Title;
1033 - Title-host-item;
1018 - Name-publisher;
5 - Title-series.

Для каждой выбранной точки доступа были определены остальные атрибуты из набора Bib-1, поддерживаемые серверами. В основном это атрибуты усечения и структуры поискового термина.

Второй справочник, помимо информации о сервере (адрес, пароль доступа), содержит информацию о каждой БД, включая шаблон запроса и поддерживаемые форматы выдачи информации.

Все это позволяет сделать вывод о том, что при создании распределенного информационного пространства даже средствами такого унифицирующего механизма как протокол Z39.50 приходится выбирать между точностью поиска и гибкостью самой системы.

Литература

1. Максимов Н.В. Компоненты и технологии интерактивного поиска документальной информации. // Международный форум по информатике и документалистике. - 2001, т.26, №3.
2. Глухов В.А., Голицына О.Л., Максимов Н.В. Электронные библиотеки. Организация, технология и средства доступа // НТИ. -Сер. 1, -2000, - №10.
3. Жижимов О.Л. Введение в Z39.50. - Новосибирск: Изд-во НГОНБ, 2000.

Об авторах

***М*аксимов Н.В.** - профессор, доктор технических наук, заведующий кафедрой "Информационных ресурсов" Российского государственного гуманитарного университета.

Сысойкина М.А. - аспирантка РГГУ, сотрудник ИНИОН РАН.
E-mail: maria@inion.ru

© *Максимов Н.В., Сысойкина М.А., 2002*