РОССИЙСКИЙ НАУЧНЫЙ ЭЛЕКТРОННЫЙ ЖУРНАЛ
Электронные библиотеки
Издается с 1998 года
ISSN 1562-5419

Электронные библиотеки -1999 - Том 2 - Выпуск 4

Метаданные и первые результаты каталогизации Интернет

М.Е. Шварцман
Российская государственная библиотека

В 1998 в Российской государственной библиотеке (РГБ) при поддержке Российского фонда фундаментальных исследований (РФФИ) началась работа по созданию систематического каталога российских ресурсов Интернет. Конечной целью этого проекта было создать базу данных (БД) описаний ресурсов российской зоны Интернет. БД должна быть доступна через Интернет, в ней должна быть предусмотрена возможность поиска по всем элементам описания ресурса и по индексам Библиотечно-библиографической классификации (ББК).

Проект рассчитан на три года, и к концу этого периода система должна быть устроена так, чтобы она смогла функционировать самостоятельно - без поддержки РФФИ. По истечении 1 года работы нам уже есть что представить российскому Интернет-сообществу.

Нами организован сайт, посвященный этому проекту http://www.rsl.ru/dc, выставлен каталог с первой тысячей описаний, появилось понимание проблем и путей их решения.

Полнота каталога

Для определения полноты каталога необходимо было понять, какие ресурсы нужно каталогизировать и как их найти. Для начала мы ограничили тематику фундаментальными науками. Для поиска ресурсов мы использовали поисковые средства RAMBLER и YANDEX. В поисковой строке задавались понятия, относящиеся к выбранной области, взятые из словесных формулировок таблицы ББК. При просмотре ресурсов нам в большом количестве встречались рекламные объявления. Предлагали свои услуги преподаватели математики и физики, предлагались различные забавные задачи по математике для начальной школы и многое другое.

Мы приняли решение отбирать только те ресурсы, которые будут интересны ученым или студентам. Оценки были, конечно, субъективные, и многое зависело от каталогизатора, но ничего лучшего нам придумать не удалось.

Также субъективно выбирался и уровень детализации при описании ресурса. Некоторые домены второго уровня, посвященные узкой теме, например http://www.akin.ru (акустический институт) каталогизировались целиком, а другие - например, сервер физфака МГУ - расписывались более подробно.

Какой формат описания выбрать

Для создания каталога необходимо выбрать стандарт описания электронных ресурсов Интернет. Для описания ресурсов Интернет можно использовать формат MARC (используемый Библиотекой Конгресса США и рядом других библиотек), тем более что разработано специальное расширение MARC для электронных документов. Этот формат позволяет очень детально каталогизировать электронный документ аналогично традиционной книге. Однако подобная детализация затрудняет использование MARC без соответствующего обучения и недоступно широкому кругу пользователей, создающих информационные ресурсы в Интернет.

В связи с этим мировым сообществом были выработаны рекомендации по набору полей и методам каталогизации информационных ресурсов, достаточно простым и доступным для их создателей. В этот набор, названный по имени семинара, где он был выработан, Dublin Core Metadata Set (DC) (http://purl.oclc.org/dc/), входит 15 полей, в которых описываются основные характеристики информационного ресурса. Поля могут повторяться и, кроме этого, поле может разбиваться на подполя. Подполе - это информация, уточняющая информацию поля. В настоящее время стандартизирован только набор полей DC. Ряд полей имеет подполя, перечень которых еще не полностью определен. Этим занимается специальная рабочая группа, организованная на пятом семинаре DC. Формат DC можно использовать как для сложного описания, используя все возможные подполя, так и для простого, обходясь без подполей вовсе. Правда в последнем случае качество описания будет хуже и вероятность нахождения потребителем данного ресурса соответственно будет

меньше.

Данный стандарт DC полностью соответствует HTML и поэтому описание ресурса, сделанное в этом формате, может быть непосредственно включено в сам ресурс при помощи меток <META>. Нужно сказать, что такая возможность, на наш взгляд, особенно важна для каталогизации ресурсов Интернет. Если ресурс содержит свое описание внутри себя, то он легко идентифицируется. В противном случае, если в базе данных существует только ссылка на URL ресурса, то при всех передвижениях ресурса найти его будет невозможно. Мета-описание внутри ресурса аналогично каталогизации в издании.

Основное отличие нашего подхода в том, что описание ресурса, размеченное по правилам HTML, будет находиться в самом ресурсе. Далее любой робот, собирающий сведения о ресурсах Интернет и понимающий наши правила описания, всегда будет иметь самые последние сведения о ресурсах содержащих внутри себя свои собственные описания.

Исходя из вышесказанного, для каталогизации нами был выбран формат Dublin Core Metadata Set. Мы используем следующие поля и подполя:
Title
Подполя:
DC.Title - Основное заглавие (подполе по умолчанию)
DC.Title.Alternative - Альтернативное заглавие
Author or Creator Лицо или организация, ответственные за содержимое ресурса.
Подполя:
DC.Creator - автор (подполе по умолчанию)
DC.Creator.PersonalName - имя индивидуального автора
DC.Creator.CorporateName - имя коллективного автора
DC.Creator.PersonalName.Address - адрес индивидуального автора
DC.Creator.CorporateName.Address - адрес коллективного автора
Subject and Keywords - предмет и ключевые слова
Description - описание
Publisher - издатель
Подполя:
DC.Publisher - издатель (подполе по умолчанию)
DC. Publisher.PersonalName - имя издателя (лица)
DC. Publisher.CorporateName - наименование издающей организации
DC. Publisher.PersonalName.Address - адрес издателя (лица)
DC. Publisher.CorporateName.Address - адрес издающей организации
Other Contributor - сведения об ответственности
Подполя:
DC.Contributor - лицо или организация, участвовавшие в создании ресурса (подполе по умолчанию)
DC.Contributor.PersonalName - имя лица, участвовавшего в создании ресурса
DC.Contributor.CorporateName - наименование коллектива, участвовавшего в создании ресурса
DC.Contributor.PersonalName.Address - адрес лица, участвовавшего в создании ресурса
DC.Contributor.CorporateName.Address - адрес коллектива, участвовавшего в создании ресурса
Date - дата
Подполя:
DC.Date.Creation_of_intellectual_content - дата создания ресурса
DC.Date.Creation/Modification_of_present_form - дата создания/модификации в существующем виде
DC.Date.Formal_publication - дата издания
DC.Date.Available - дата предоставления в открытом доступе
DC.Date.Valid (includes verification) - дата, начиная с которой информация в ресурсе будет соответствовать действительности
DC.Date.Acquisition/Accession - дата поступления материалов
DC.Date.Accepted - дата приобретения материалов
DC.Date.DataGathering - дата сбора материалов
Resource Type - тип ресурса
Format - формат
Resource Identifier - идентификатор ресурса
Source - источник
Language - язык
Relation - отношения
Подполя:
DC.Relation.Creative - первоисточник (для переводов и аннотаций)
DC.Relation.Mechanical - оригинал (в случае зеркального копирования, или копирования без изменения содержания)
DC.Relation.Version - версия (указываются предыдущие редакции ресурса)
DC.Relation.Inclusion - составная часть (адрес ресурса, составной частью которого является описываемый ресурс или составная часть описываемого ресурса
DC.Relation.Reference - ресурс, на который делается ссылка или из которого берутся цитаты.
Coverage - зона действия или охвата

Подполя:
DC.Coverage.PeriodName - охватываемый период времени
DC.Coverage.PlaceName - охватываемые территории
DC.Coverage.t - координата
DC.Coverage.x - координата
DC.Coverage.y - координата
DC.Coverage.z - координата
DC.Coverage.Polygon - описание двумерного объекта
DC.Coverage.Line - описание линейного объекта
DC.Coverage.3d - описание трехмерного объекта
Rights Management - правовые аспекты.

Проблемы с каталогизацией

При каталогизации многих ресурсов было очень трудно найти их формальные характеристики. На многих серверах отсутствуют имена их создателей. Может присутствовать фамилия одного web-мастера и остается только гадать, только ли он оформлял сервер или же создавал его содержание. Часто отсутствует дата издания. Порой трудно выделить заглавие из художественных изысков дизайнера. Ряд полей, предусмотренных нами, сейчас практически не используется. Большое разнообразие дат и разработанность поля "Охват" остается невостребованным. В описываемых нами ресурсах нет информации для заполнения этих полей.

В общем, трудности описания аналогичны трудностям каталогизатора печатного издания, а может быть даже большие. Если у издателей печатной продукции есть какие-то стандарты или хотя бы традиции, то при создании Homepage каждый web-мастер творит в полную силу своей фантазии

Создание программного обеспечения для ведения БД

В начале предполагалось, что база данных будет создаваться силами участников проекта. Так сейчас и происходит: все записи в БД введены нами. Однако в дальнейшем необходимо участие всех создателей ресурсов в формировании этого каталога. Мы исходим из того, что создатели ресурсов более других заинтересованы в расширении круга посетителей своего ресурса. Следовательно, они должны быть готовы создать описание своего ресурса-им нужно только дать инструмент для создания описания.

Соответственно, мы делали программное обеспечение, пригодное для использования создателем ресурса. При помощи этого инструмента каждый желающий может получить описание своего ресурса в формате DC. Наше программное обеспечение (ПО) позволяет вводить описание ресурса через заполнение HTML-формы в окне броузера. После заполнения формы ПО возвращает пользователю описание его ресурса в формате DC. Это описание может быть вставлено при помощи любого текстового редактора между метками <HEAD> </HEAD>

Примеры описания в формате Dublin Core начальной страницы Web-сайта Российской государственной библиотеки, возвращаемого после заполнения формы.

<META NAME="DC.Title" CONTENT="Российская государственная библиотека">
<META NAME="DC.Creator.PersonalName" CONTENT="Власенко Т.В.">
<META NAME="DC.Creator.CorporateName" CONTENT="Отдел автоматизации РГБ" >
<META NAME="DC.Creator.PersonalName.Address" CONTENT="vlas@rsl.ru" >
<META NAME="DC.Creator.CorporateName.Address" CONTENT="noa@rsl.ru" >
<META NAME="DC.Subject" SCHEME="ББК" CONTENT="78.34(2)">
<META NAME="DC.Description" CONTENT="Представлены история и современное состояние Российской государственной библиотеки. Перечислены информационные ресурсы и в том числе доступные через Интернет">
<META NAME="DC. Publisher.CorporateName" CONTENT="Российская государственная библиотека">
<META NAME="DC. Publisher.CorporateName.Address" CONTENT="101000, Москва, Воздвиженка, 3">
<META NAME="DC.Contributor.PersonalName" CONTENT="Козлова Н.В., Web-мастер" >
<META NAME="DC.Contributor.PersonalName.Address" CONTENT="webmaster@rsl.ru" >
<META NAME="DC.Date.Creation_of_intellectual_content" CONTENT="1997-06-01">
<META NAME="DC.Date.Creation/Modification_of_present_form"
CONTENT="1998-01-01">
<META NAME="DC.Date.Available" CONTENT="1997-06-01">
<META NAME="DC.Type" CONTENT="Homepage.Organization">
<META NAME="DC.Identifier" CONTENT="www.rsl.ru">
<META NAME="DC.Language" SCHEME="ISO 639-2" CONTENT="rus">
<META NAME="DC.Rights" CONTENT="Свободный доступ">

Для ведения БД был выбран Paradox for Windows. Основной причиной выбора данной СУБД была та, что это в настоящий момент единственное официально купленное ПО для ведения БД. Пока БД небольшая, Paradox нас удовлетворяет, но позднее, несомненно, придется перейти на более мощную СУБД.

Для передачи информации из HTML-формы в БД были написаны CGI программы, а для проведения формально-логического контроля написаны крипты на Java.

Большой сложностью было создание механизма повторяющихся полей в HTML-форме. Статические HTML страницы для этого не годились, так как неизвестно, сколько повторов поля потребуется для описания ресурса. Для решения этой проблемы мы создали механизм динамического создания страниц с помощью специально программы, изменяющей количество полей в форме по мере необходимости. В наших ближайших планах написание программы для преобразования метаданных в формате DC в базу данных. Это позволит в дальнейшем создавать роботов для просмотра Интернет и для выборки ресурсов содержащих метаданные и записи их в БД

Проблемы дублетности ресурсов

В настоящее время единственным идентификатором ресурса является его URL (Universal Resource Locator).

При каталогизации ресурсов мы столкнулись с тем, что один и тот же ресурс может находиться в разных местах (зеркало) или в том же самом месте, но в другой кодировке. URL у таких ресурсов разный и, формально подходя, их нужно описывать еще раз. Но с другой стороны - это ведь один и тот же ресурс. Если ресурс переезжает с одного сервера на другой и меняет URL, он также не становится от этого другим ресурсом.

Аналогично, каждая книга в библиотеке имеет свой "URL" - шифр хранения. Однако она также имеет и ISBN, однозначно идентифицирующий издание.

Мировое сообщество осознало необходимость внедрения аналогичной идентификации ресурсов Интернет и сейчас разрабатывается система URN (Universal Resource Number). Более подробно об URN рассказывается в специальном докладе. Пока такая система не внедрена, мы решили при описании разных копий одного ресурса выбирать одну из копий за основную, а URL остальных указывать как место хранения копий.

Проблема поддержания актуальности каталога

Даже за то небольшое время, что мы создаем свой каталог, некоторые из описанных нами ресурсов стали недоступными. Причины этого могут быть разными: финансовые проблемы владельца сервера, аварии линии связи и т.п. В результате каталог становится не актуальным. Для поддержания каталога в актуальном состоянии возможны два подхода:

- регулярно проверять все ссылки в БД и удалять описание ресурсов, к которым нет доступа;
- перекачивать в БД содержание ресурса и не зависеть от капризов судьбы.

Первый подход проще в реализации, но очень жалко потерянного времени на каталогизацию потерянных ресурсов.

Второй подход требует значительных затрат, но зато создает архив ресурсов - аналог Книжной палаты. Такой инструмент необходим для изучения культурного наследия страны. В рамках нашего проекта у нас не хватает средств на реализацию второго подхода. Однако мы пробуем осуществить его для полных текстов художественных произведений. Более подробно см. описание проекта OREL (Open Russian Electronic Library) http://orel.rsl.ru.

Проблема поиска новых ресурсов

Когда мы только приступили к созданию каталога, каждый найденный ресурс мог быть введен в БД без проверки на дублетность. С ростом БД все больший процент найденных ресурсов оказывается уже описанным. Как показал опыт, большая часть времени уходит не на каталогизацию, а на проверку дублетности.

В настоящее время мы автоматизировали этот процесс. Все ссылки, найденные поисковыми системами, сохраняются в HTML-файле. Затем наша программа в пакетном режиме просматривает его и проверяет в базе наличие записей с такими же URL. Отсутствующие в БД помечаются и затем каталогизируются.

Однако более перспективным направлением представляется привлечение широкого круга создателей ресурсов и процессу каталогизации своих произведений. Необходимо, чтобы создание метаданных стало элементом культуры любого Web-мастера.

При наличии такого описания возможно создание программы-робота, которая без участия человека будет просматривать весь Интернет или его часть, находить новые или измененные ресурсы, выбирать из ресурсов описания, подготовленные создателями и размещать их в создаваемом каталоге. Таким образом, обеспечивается актуальность каталога, а его полнота будет зависеть от создателей ресурсов.

Ссылки

1. Шварцман М.Е. К вопросу каталогизации ресурсов Интернет // Мир библиографии. - N 5, 1998.
2. Шварцман М.Е. Использование метаданных для каталогизации российских ресурсов Интернет // Электронные библиотеки. Т.1, Вып. 2, 1998 - URL: http://www.elbib//1998/199802/Shvarz/shvarz.ru.html [20 мая 1999]
3. Семинар по метаданным в Хельсинки: Отчет о семинаре и последующих разработках / Стюарт Вебель, Юха Хакала. // Электронные библиотеки. - Т. 1, Вып. 2, 1998 - URL: http://www.elbib//199802/WH/wh.ru.html [20 мая 1999]
