ОРДЕНА ЛЕНИНА
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
им. М.В.Келдыша
Российской Академии Наук

Е.Л.Китаев, Д.Л.Кузьмичев, М.И.Слепенков

Проблемы технологического обеспечения многоцелевого режима эксплуатации каталогов метаданных Интернет

Москва, 2002
Настоящая работа поддержана Российским фондом фундаментальных исследований, грант № 02-01-00283

Аннотация

В настоящей работе делается попытка представить обзор современной проблематики технологического обеспечения каталогов метаданных Интернет. На примере каталога веб-сайтов дается описание архитектуры системы, вводятся основные сущности и структуры данных, в которые они отображаются, а также рассматривается процесс ведения каталога, компоненты и рабочие места, участвующие в этом процессе. Приводятся основные функциональные требования к системе обслуживания подписчиков при эксплуатации каталога в многоцелевом режиме, а также предлагаются некоторые решения, направленные на удовлетворение этих требований.

Technological problems for supporting multi-purpose usage of Internet metadata catalogues

Abstract

The paper attempts to give an overview of current technological problems for supporting Internet metadata catalogues. Based on the examples of a Web-site catalogue authors present the architecture, introduce underlying objects and data structures used for their representation in the database. The process of the catalogue content management is described along with the components and tools participating in the process. The set of functional requirements for catalogue subscriber service subsystem is discussed and some solutions are proposed.

СОДЕРЖАНИЕ

Введение

Сегодня одной из самых распространенных задач, которую приходится решать пользователям глобальной сети, является поиск и оптимизация доступа к ресурсам Интернет. Для обеспечения этой потребности в сети создаются узлы-посредники, в основе организации которых лежат массивы метаданных - определенным образом структурированная информация, описывающая атрибуты ресурсов. Эффективность работы посредника напрямую зависит от качества метаданных, которыми он располагает. При этом важную роль играет не только формальный аспект (например, степень структурированности метаданных), но и достоверность самой информации о ресурсе, ее актуальность.

На сегодняшний день высокое качество описания ресурсов достигается в каталогах метаданных. Каталоги обычно специализируются на определенном типе ресурсов: описания сайтов, электронных документов, дистрибутивах программ, базах данных, вычислительных серверах сети и т.п. В каталогах, автоматизированные процедуры сбора информации (сканирование сети, организация приема заявок от владельцев ресурсов и т.п.) дополняются участием экспертов (референтов). В задачи экспертов входит оценка различных параметров ресурса, его классификация, составление описания ресурса и задание ряда атрибутов, которые невозможно определить автоматически.

Особенностью метаданных ресурсов Интернет, отличающих их от прочих информационных фондов (библиографических каталогов, музейных каталогов и т.п.), является высокая степень изменчивости предмета описания. Помимо изменения содержания самого ресурса(например, обновление контента сайта, изменения состояния загруженности вычислителя), объективно существует такой важный параметр – как доступность. Понятно, что для поддержки

актуальности метаданных процесс экспертизы ресурса должен производиться итеративно и достаточно регулярно. Из вышесказанного следует, что организация каталога метаданных ресурсов Интернет является дорогостоящим и трудоемким предприятием.

Одним из перспективных путей обеспечения подписчиков качественной информацией при одновременном снижении ее себестоимости является переход от одно-целевого режима функционирования каталогов метаданных (т.е. когда каталог работает в качестве поставщика информации для одного потребителя), к много-целевому режиму, когда каталог обслуживает большую группу подписчиков - потребителей метаинформации -одновременно. Основной проблемой, возникающей при таком варианте использования каталога, является многообразие взглядов на метаданные со стороны подписчиков. Такая постановка объективно вытекает из того, что на практике подписчики специализируются не только по типам ресурсов, но и по целевым группам пользователей, запросы которых они обслуживают.

В настоящей работе делается попытка представить обзор современной проблематики технологического обеспечения каталогов метаданных Интернет. Содержание работы разделено на два раздела. В первом разделе будет описана архитектура каталога, введены основные сущности и структуры данных, в которые они отображаются, а также рассмотрен процесс ведения каталога и компоненты, участвующие в этом процессе.

Во втором разделе будут описаны основные функциональные требования к системе обслуживания подписчиков, возникающие при эксплуатации каталога в многоцелевом режиме, а также предложены некоторые решения, обеспечивающие выполнение этих требований. В заключительной части можно найти краткий обзор наиболее значимых инициатив в области организации инфраструктуры метаданных Интернет и месте настоящей работы в этой научно-практической области.

1 Архитектура каталога метаданных

Цель этого раздела - сформировать у читателя общее представление об архитектуре каталога метаданных, а также рассмотреть широкий круг технических деталей, связанных со структурами данных, в которые погружается информация; технологические операции по ведению контента; принципы организации редакционного процесса; связи с внешними системами и сервисами и т.д. Последующее изложение будет построено на примере каталога описаний Веб-сайтов. Авторы исходят из того, что подавляющее большинство читателей знакомо с каталогами этого класса, составляющих основу популярных порталов www.yahoo.com, www.yandex.ru, www.aport.ru, и поэтому нет необходимости раскрывать целевую функцию этого типа метаданных.

1.1 Основные сущности каталога

В основе любого каталога лежат три сущности: рубрикаторы, состоящие из рубрик (тематических разделов, категорий, географических понятий и т.п.) образующих иерархическую структуру; объекты рассматриваемой области – описания ресурсов, характеризующиеся определенным набором атрибутов; а также связи, которые устанавливаются между ресурсами и рубриками (позиционирование ресурсов в рубрикаторе).

Все обладающие общими свойствами объекты можно выделить в группу. Объекты внутри этой группы можно разделить на подгруппы, определяемые другими свойствами, еще более конкретизирующими объект. Таким образом, можно выделять подгруппы вплоть до полного исчерпания известных свойств объектов или до нужного уровня конкретизации (абстракции). Рубрика – это некий уровень абстракции, классифицирующий объект, а совокупность рубрик образует иерархическую систему и называется рубрикатором. На Рис. 1 приведена примеры простой иерархической структуры тематического и географического рубрикаторов Веб-сайтов.

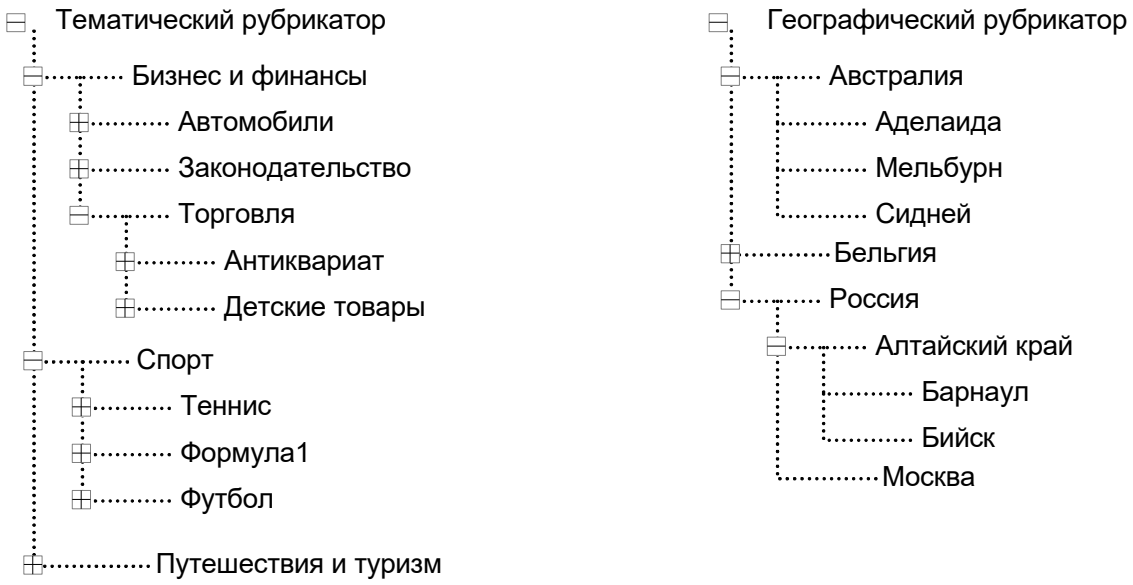

Рис. 1 Фрагменты тематического и географического рубрикаторов

Как видно из рисунка, на самом верхнем уровне иерархии находится абстрактный раздел с названием рубрикатора. Он содержит, например, такие подрубрики, как "Бизнес и финансы", "Спорт", "Путешествия и Туризм". Те, в свою очередь, содержат другие подрубрики и т.д.

Некоторые рубрики могут быть по смыслу ассоциативно связаны с другими группами рубрик, находящимися на других ветвях иерархии. Пример подобных ассоциаций приведен на рис. 2. Как видно из рисунка, такими связями могут обладать подразделы "Бизнес" и "Спорт" раздела "СМИ", которые образуют смысловые связи с подразделами "Новости бизнеса"

(раздел "Бизнес и финансы") и "Новости спорта" (раздел "Спорт"). На рис. 2 эти связи указаны в виде ссылок на соответствующие разделы "СМИ".

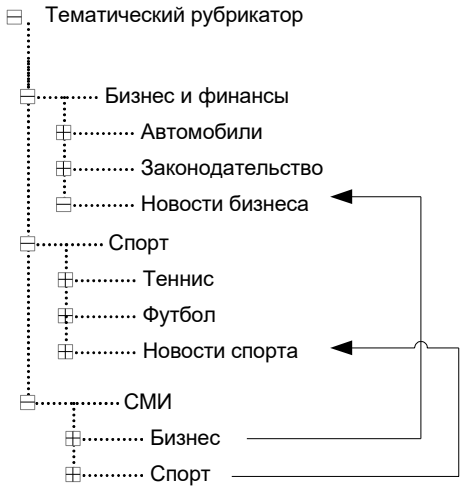

Рис 2. Пример ассоциативных связей между рубриками

Другими словами, все объекты, принадлежащие, например, разделу "СМИ/Бизнес", могут быть отнесены в виртуальный подраздел "Новости бизнеса" раздела "Бизнес и финансы".

Основные атрибуты, которыми должны обладать ассоциативные связи – это идентификатор раздела, содержащий объекты; идентификатор раздела, в котором будет создан виртуальный подраздел, и псевдоним (название) виртуального подраздела.

Переходя к атрибутному составу описания Веб-сайта, следует сразу отметить, что оно не такое простое, как может показаться посетителю страниц каталога. Данное описание отнюдь не исчерпывается URL сайта и его аннотацией. Ниже представлена форма – регистрационная карточка описания Веб-сайта, которая используется при ведении действующего каталога www.virtalog.ru.

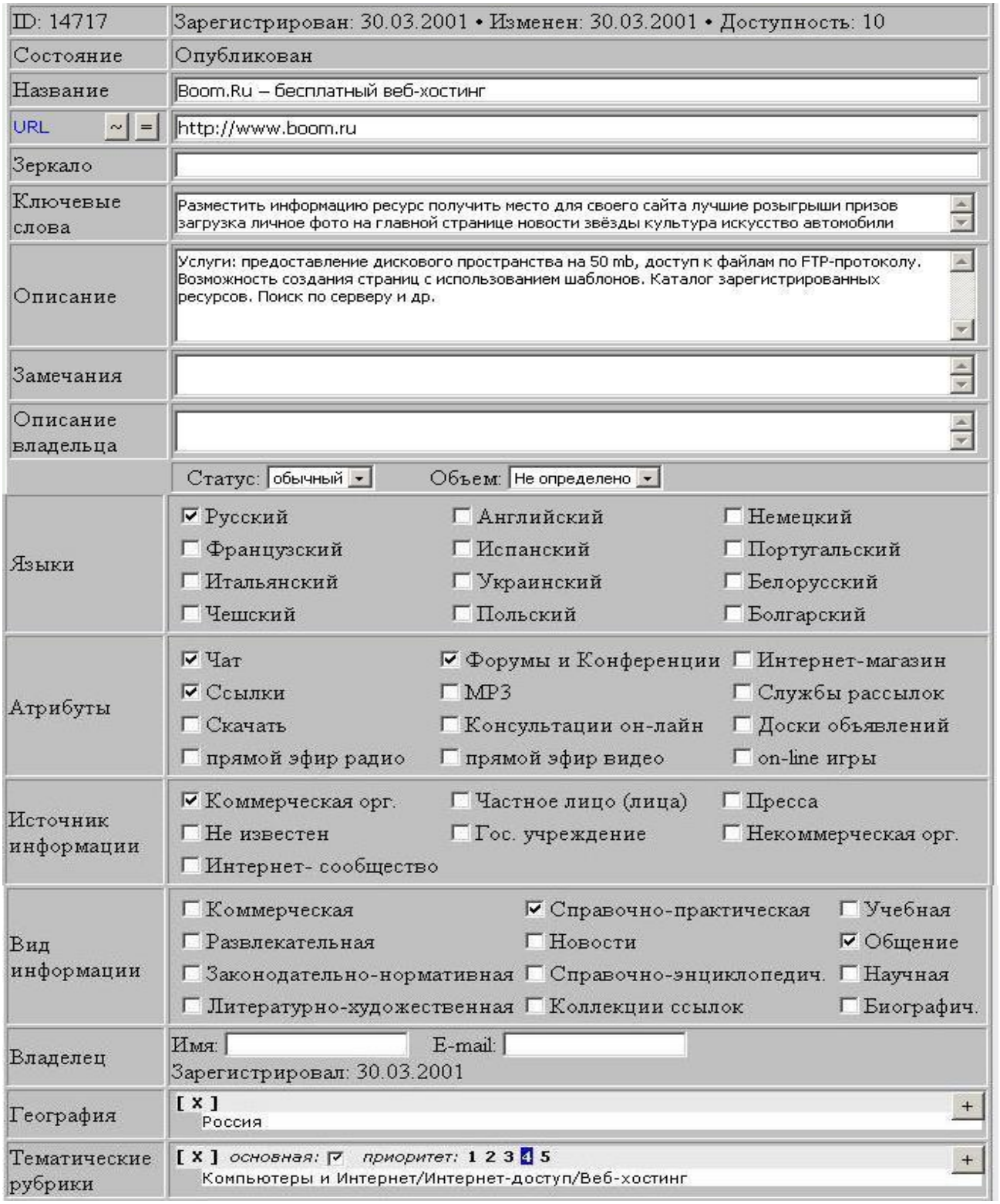

Рис. 3 Карточка описания веб-сайта

Полный перечень реквизитов описания ресурса включает в себя:

- URL ресурса – адрес сайта.

- Заголовок содержит название и цель сайта, так, чтобы читая заголовок отдельно от аннотации, посетитель мог понять основную идею сайта.
- Аннотация сайта – краткое описание содержания сайта, не содержащее оценочной или рекламной информации.
- Аннотация от владельца сайта – дополнительная информация об изменениях и дополнениях к содержанию сайта.
- Ключевые слова – набор слов, по которым может быть осущёствлен поиск сайта, дополняющий слова из заголовка или аннотации.
- Язык сайта – множественный признак, определяющий на каких языках представлены публикации на сайте.
- Сервисы сайта – перечень признаков, отражающих наличие на сайте специализированных информационных разделов или функций (чата, конференций и форумов, подборок ссылок, интернет-магазина, службы рассылок, MP3 и др.).
- Источник информации (Коммерческая организация, Частное лицо, Пресса, Неизвестен и т.п.) – дополнительный признак, позволяющий посетителю точнее оценить уровень достоверности информации, представленной на сайте.
- Вид информации (Коммерческая, Развлекательная, Справочная, Общение и т.п.) – дополнительный признак, позволяющий посетителю оценить характер представленной на сайте информации.
- География – элементы специального географического рубрикатора, характеризующие региональную принадлежность содержимого сайта, могут отсутствовать у ресурсов, не имеющих по своей природе региональной принадлежности.
- Тематические рубрики – элементы тематического рубрикатора, отражающие тематическую направленность ресурса. Для каждой рубрики редактором дополнительно устанавливается коэффициент релевантности ресурса и темы. Этот коэффициент обычно используется, как один из параметров сортировки при показе списка ресурсов в рубрике на страницах каталога.
- Объем ресурса – атрибут принимает одно из следующих значений: Страница, Сайт, Раздел сайта, Не определено.
- Доступность ресурса – результат работы специальной утилиты, осуществляющей регулярную автоматическую проверку доступности ресурса по его URL.
- Служебные реквизиты – идентификатор ресурса в БД; реквизиты, характеризующие состояние обработки ресурса (зарегистрирован, в работе, в архиве, опубликован и т.п.), дату регистрации в каталоге, дату актуализации и т.п.

Как видно, структура описания ресурса является достаточно сложной с большим количеством множественных атрибутов, а также связями с рубрикаторами (при этом каждая связь обладает собственными атрибутами). Важно также отметить, что в описании ресурсов представлены атрибуты, которые имеют различный регламент ведения: автоматический (например, доступность); полуавтоматический (например, название и ключевые слова могут быть получены путем сканирования головной страницы сайта, а затем обработаны редактором); большая часть полей ведется вручную параллельно редакторским коллективом каталога и владельцами сайтов (в частности, атрибуты сайта, указанные его владельцем, могут быть проверены и скорректированы редактором).

1.2 Представление основных сущностей каталога в базе данных

Определив три основных сущности каталога и его атрибуты, мы тем самым определили несколько таблиц базы данных и отношения между ними. Типовая схема базы данных Веб-каталога будет иметь вид, представленный на Рис. 4.

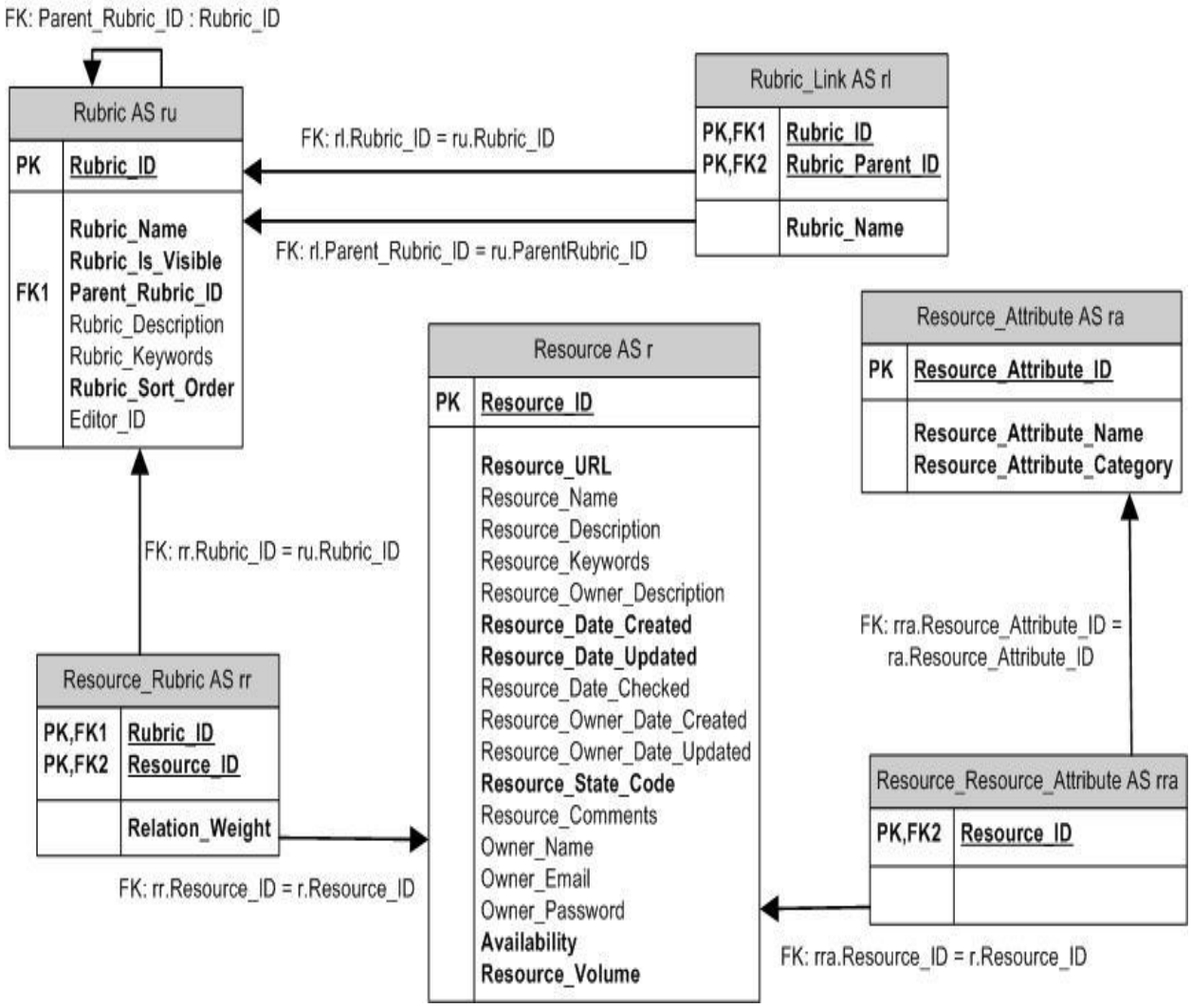

Рис. 4 Фрагмент схемы Базы Данных для представления основных сущностей каталога

В схеме представлено шесть таблиц:

- **Resource**– таблица, содержащая описания сайтов: Resource_ID – идентификатор сайта в БД; Resource _URL – адрес сайта в Интернет; Resource_Name – название сайта; Resource_Description – описание сайта, публикуемое на страницах каталога (смысл остальных полей мы не рассматриваем, поскольку он очевиден из их названий и соответствует тем атрибутам, которые представлены в карточке описания сайта см. Рис.3).
- **Rubric** – таблица для представления иерархических рубрикаторов Веб-каталога: Rubric_ID – идентификатор рубрики - тематического раздела каталога; Rubric_Parent_ID – идентификатор родительского раздела; Rubric_Name – название раздела; Rubric_Description – описание раздела; Rubric_Keywords – список ключевых слов характеризующих рубрику; Rubric_Sort_Order – поле, значение которого используется для упорядочивания подрубрик на страницах Веб-каталога при показе

тематического раздела и всех его подрубрик; Rubric_Is_Visible – поле типа bit, принимающее одно из двух значений: 1 - рубрика публикуется на страницах каталога, 0 – рубрика не публикуется (обычно «невидимые» рубрики используются в технологических целях для организации редакторского процесса, о котором речь пойдет ниже); Editor_ID – идентификатор редактора, отвечающего за ведение ресурсов в данном разделе каталога.

- **Resource_Rubric** – таблица для представления множественных связей между веб-ресурсами и рубриками классификатора: Resource_ID – идентификатор сайта в БД; Rubric_ID – идентификатор рубрики, к которому принадлежит сайт; Relation_Weight – весовой коэффициент, значение которого используется для сортировки описаний сайтов на страницах Веб-каталога при показе всех ресурсов, связанных с данной рубрикой.
- **Rubric_Link** – таблица, содержащая описание ассоциативных связей между рубриками: Rubric_ID – идентификатор рубрики; Rubric_Parent_ID – идентификатор дополнительного родительского узла; Rubric_Name – имя псевдонима рубрики, с которым она будет показана в контексте данного родительского узла.
- **Resource_Attribute** – таблица с нормативными данными для представления атрибутов сайта: Resource_Attribute_ID - идентификатор атрибута; Resource_Attribute_Name - имя атрибута; Resource_Attribute_Category – категория, к которой принадлежит данный атрибут (в каталоге выделяются следующие категории: Языки сайта, Сервисы сайта, Источник информации, Вид информации).
- **Resource_Resource_Attribute** – таблица, содержащая информации об атрибутах сайта: Resource_ID – идентификатор ресурса; Resource_Attribute_ID – идентификатор атрибута.

Представленный фрагмент схемы БД является одним из наиболее простых и удобных вариантов проектирования каталога, хотя существуют и другие не менее оригинальные методы для представления иерархических данных. В принципе, с помощью этих структур можно разрабатывать информационные системы различной степени сложности.

1.3 Ведение каталога

Каталог веб-сайтов является публичным ресурсом Интернет. Более того, многие каталоги поддерживают различного рода сервисы, связанные с распространением данных своим подписчикам. Сказанное означает, что к качеству описания ресурсов, представленных в каталоге, предъявляются повышенные требования. Для достижения поставленной цели образуется редакторский коллектив, в обязанности которого входит поиск и подготовка материалов для их публикации на страницах каталога. Процесс обработки

описаний веб-сайтов является итеративным, и редактор обязан постоянно отслеживать все существенные изменения происходящие с ресурсами и вносить соответствующие исправления в их описания.

В компаниях, бизнес которых связан с ведением и публикацией каталогов, для автоматизации редакторского процесса создаются специализированные рабочие места – АРМ (Автоматизированное Рабочее Место). Как правило типовой АРМ обеспечивает выполнение следующих функций:

- Ввод новых описаний веб-сайтов и сохранение их в БД. Для этого редактор заполняет карточку с описанием аналогичную той, что представлена на Рис. 3, а также классифицирует ресурс, связывая его с тематическими и географическими рубриками.
- Поиск ранее введенных описаний с целью их просмотра и коррекции. К числу атрибутов, по которым выполняется поиск, можно отнести: http-адрес ресурса (URL); его идентификатор в БД; состояние обработки описания веб-сайта; рубрики (географические и тематические); ключевые слова в названии, описании веб-сайта; дата создания и обновления описания. В результате выполнения запроса на поиск редактор получает порционный список релевантных записей. Каждый элемент списка представляет собой краткое сведение о веб-сайте (URL, название, аннотация, состояние обработки). Редактор может выбрать из списка любое описание (раскрыть элемент списка) и перейти к редактированию.
- Удаление описаний из БД. Обычно эта операция выполняется по отношению к веб-сайтам, которые прекратили свое существование, перестали отвечать тематике (или требованиям) каталога.
- Помещение описания в Архив. Архивирование описаний производится на определенный срок, когда действующий веб-сайт становится временно недоступным, например, при смене хостинговой площадки или по каким-то другим причинам.
- Ведение древовидного рубрикатора. Как правило, данная функция возлагается на главного редактора, который отвечает за пополнение рубрикатора, его реструктуризацию, отслеживание сбалансированности дерева и позиционированных в нем описаний. К числу операций связанных с ведением древовидного рубрикатора относятся: создание рубрик, установление иерархических и ассоциативных связей между ними; перемещение поддеревьев (смена родителя); редактирование названий и других атрибутов рубрик; удаление неактуальных узлов (выполнение этой операции возможно только в том случае, если с рубрикой и всеми ее подрубриками вплоть до листовых элементов не связано не одного ресурса).
- Поддержка рабочих очередей для редакторов. В обязанности главного и старших редакторов входит подготовка «пула работ» для младшего персонала. Для этого в АРМ должна быть предусмотрена функция – передать описание веб-сайта на обработку. Одним из способов реализации

данной возможности является заведение служебных непубликуемых рубрик (названиями которых являются фамилии редакторов) в классификаторе, с которыми старшие редактора связывают «сырые» описания (обычно оно содержит только URL веб-сайта). Кроме того, в классификаторе обычно заводится еще одна служебная рубрика – Очередь заявок внешних пользователей, куда попадают заявки от самих хозяев интернет-сайтов (сюда же попадают заявки, созданные автоматически, программой, сканирующей Интернет).

- Получение разнообразной статистики, отражающей как состояние каталога в целом (например, распределение описаний веб-сайтов по состояниям), так и деятельности отдельного редактора (например, выдача гистограмм по количеству обработанных описаний сотрудниками во временном разрезе и т.п.).

В процессе обработки описания веб-сайтов проходят сложный технологический цикл, который отражается в смене их состояний. Так, например, сначала описание веб-сайта регистрируется в очереди редактора в виде заявки. Поступившая заявка обрабатывается редактором, причем в процессе работы она может неоднократно сохраняться в БД и принимать состояние «в работе». Завершив составление описания, редактор может его опубликовать или, если у сотрудника нет прав на публикацию, отправить обработанный ресурс на одобрение своего непосредственного начальника. В последнем случае, старший редактор проверяет корректность составленного описания и либо публикует его, либо отправляет на доработку и т.д.

Ориентировочный список состояний может выглядеть следующим образом:

Заявка на включение в каталог – заявка находящаяся в очереди редактора или общей очереди;

В работе – описание веб-сайта обрабатывается редактором (готовится к публикации);

Опубликовано - описание веб-сайта составлено и может публиковаться на страницах каталога;

В архиве - описание веб-сайта отправлено в архив на определенный срок, по истечении которого описание возвращается в очередь редактора или общую очередь;

Из архива - описание веб-сайта, поступившее в очередь редактора или в общую очередь из архива;

Запрещен – публикация описания веб-сайта запрещено на страницах каталога. Обычно в это состояние переводятся описания веб-сайтов, тематика которых противоречит существующему законодательству и общепринятым нормам морали (такие описания умышленно не удаляются из каталога, чтобы предотвратить их повторную регистрацию);

Удален – описание веб-сайта помечено как удаленное, физическое удаление выполняется обычно хранимой процедурой чуть позже по истечению

некоторого срока (описания находящиеся в данном состояние могут быть возвращены в работу и опубликованы)
На доработку - описание веб-сайта составлено недостаточно точно и должно быть переработано, обычно в такое состояние описание переводится старшими редакторами в процессе контроля деятельности младших сотрудников - стажеров;
Предварительно опубликован, Предварительно удален, Предварительно запрещен – эти три состояния моделируют для младших сотрудников ход редакторского процесса. Описания веб-сайтов, находящиеся в данных состояниях, требуют утверждения старшего персонала, которое может снять признак «предварительно» или отправить описание на доработку.

Существенным требованием, которое должно быть выполнено при реализации АРМ'а редактора, является обеспечение разграничение доступа по ресурсам. В рамках каталога с описаниями веб-сайтов зоной ответственности каждого редактора обычно является набор рубрик, за наполнение которых он отвечает.

С целью минимизации затрат на программирование и последующее сопровождение АРМ'ы предпочтительнее всего реализовывать в виде самостоятельных серверных веб-приложений. Работа с таким АРМ выполняется с помощью штатного броузера, входящего в поставку операционной системы, установленной на компьютере редактора. Подобное решение не требует выполнения инсталляции на рабочих местах дополнительного программного обеспечения, а также существенно облегчает внесение и распространение изменений в модули, поскольку выполняется в одном месте, на Веб-сервере.

2 Многоцелевое использование каталога метаданных

Из сказанного в предыдущем разделе несложно сделать вывод, что ведение качественного каталога метаданных является сложным и дорогостоящим процессом. Поэтому организаторы каталога заинтересованы в создании и развитии разнообразных сервисов (в том числе и платных), которые бы обеспечили возможность многоцелевого использования подготовленного контента. Успех подобных проектов в значительной степени определяется тем, насколько удобно и эффективно реализованы механизмы обслуживания широкого круга подписчиков.

2.1 Реструктуризация данных

Одна из главных проблем, возникающих при многоцелевом варианте использования каталога, является многообразие взглядов на метаданные со стороны подписчиков. Такая постановка объективно вытекает из того, что на практике подписчики специализируются не только по типам ресурсов, но и по целевым группам пользователей, запросы которых они обслуживают.

Прежде всего, каталог должен обладать механизмами, с помощью которых подписчик мог бы специфицировать свой персональный взгляд на структуру каталога. В простейших случаях это достигается за счет создания соответствующих образов (view) в структуре базы данных. С помощью этого средства могут быть реализованы функции переименования атрибутов, несложное конвертирование их значений, исключение из взгляда не интересующих подписчика атрибутов и т.п. Таким способом обслуживается большинство клиентов, которых удовлетворяет содержание каталога, но не устраивает формат представления метаданных.

Однако, со стороны подписчиков зачастую возникает потребность в изменении (дополнении) атрибутного состава в метаописании. Например, разные подписчики предъявляют свои требования к размеру текста в поле «Аннотация ресурса». Кто-то считает, что аннотация должна быть сжатой (не более 80 символов), а кто-то, наоборот, предпочитает иметь в этом поле развернутое описание. В подобных случаях, когда в структуре каталога не хватает нужных атрибутов, подписчик должен иметь возможность инициировать процедуру расширения атрибутов. В принципе, подобная процедура не может быть автоматизирована, однако архитектура каталога (его программная реализация) должна позволять «безболезненно» (в частности, для редакторского процесса) пополнять структуру метаописания.

Другой сложной ситуацией, с которой сталкиваются разработчики многоцелевого каталога, является наличие у подписчиков собственных массивов нормативных данных (рубрикаторов, словарей и т.п.), которые они хотели бы связать с метаданными каталога.

Эта ситуация может быть разделена на два случая. В первом случае, когда в структуре каталога нет соответствующих по смыслу атрибутов или рубрикаторов, задача сводится к их добавлению в существующую структуру. С этой целью в составе программного обеспечения каталога должны быть предусмотрены средства импортирования нормативных данных от подписчика и их последующей инкрементальной актуализации (если подписчик не делегирует права на ведение своих нормативных данных персоналу каталога). Во втором случае, когда в структуре каталога уже существуют аналогичные нормативные данные, возникает задача установления их отображения на нормативные данные подписчика.

Одним из вариантов решения задачи отображения (мэппинга) рубрикаторов друг на друга является создание утилиты, при помощи которой можно установить соответствие между рубриками двух рубрикаторов (рубрики исходного рубрикатора отображаются на рубрики целевого). Экранная форма утилиты представляет собой фрейм, разделяющий экран на три области (см. рис. 5). В левую часть будет загружен целевой рубрикатор, а в правую часть – исходный. Центральная область - область мэппинга - разделена на секции. Каждая секция соответствует отдельной рубрике целевого (расположенного слева) рубрикатора и начинается с заголовка, в котором представлено название рубрики. Под заголовком секции будет

располагаться тело секции, в которое можно перемещать рубрики исходного (расположенного справа) рубрикатора, связываемые с данной целевой рубрикой. После перемещения рубрик в тело секции мэппинга можно дополнительно указать уточняющие атрибуты (например, переносить только объекты, непосредственно связанные с данной рубрикой, либо все объекты, связанные с рубрикой и всеми ее подрубриками).

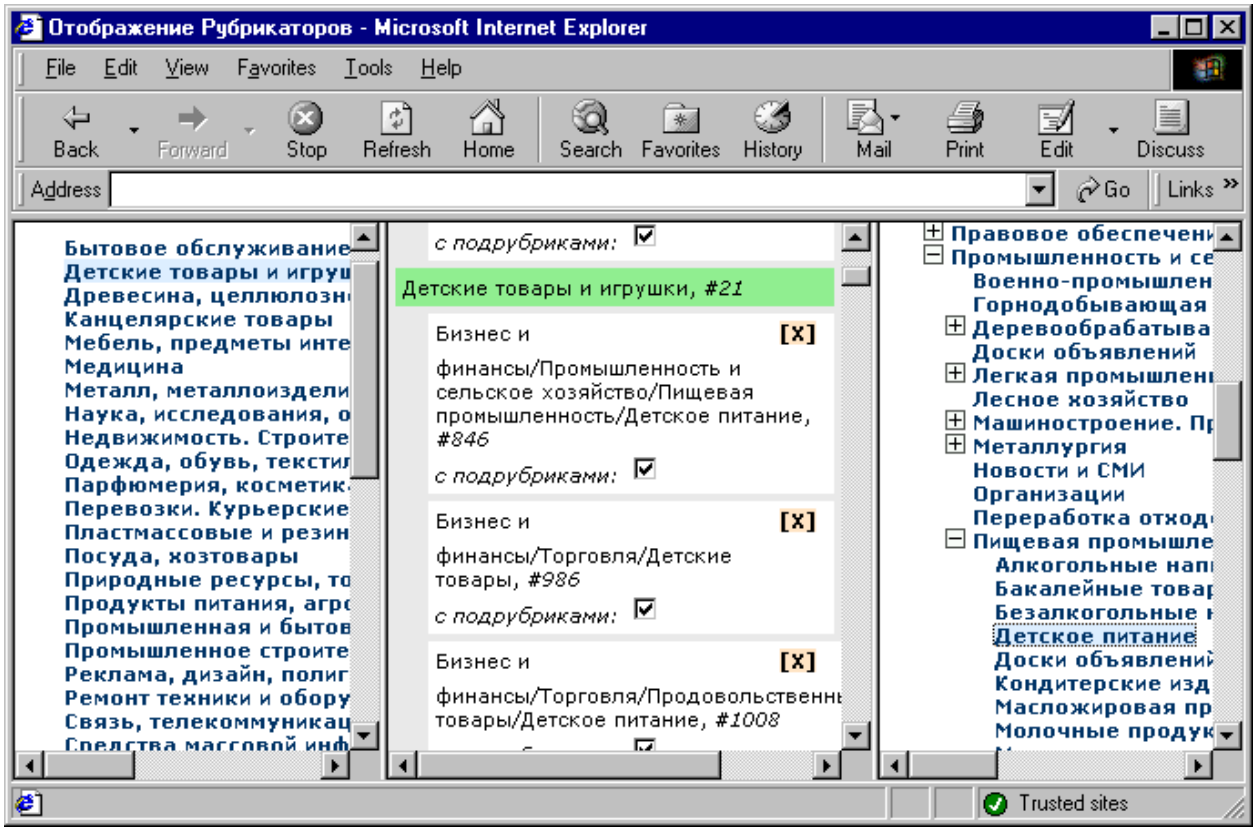

Рис. 5 Экранная форма утилиты отображения (мэппинга) рубрикаторов

На основании подобного меппинга можно автоматически отобразить связи объектов с рубриками исходного рубрикатора на связи объектов с рубриками целевого рубрикатора (т.е. выполнять автоматическую перерубрикацию). Понятно, что установление меппинга несравнимо по трудоемкости с выполнением ручной рубрикации объектов.

2.2 Фильтрация информации и интеграция с данными подписчика

Часто оказывается, что подписчика интересует не весь контент, имеющийся в каталоге, а лишь небольшая его часть. Например, значительную часть подписчиков каталога веб-ресурсов составляют веб-сайты, специализирующиеся на определенной тематике (бизнес порталы, спортивные порталы и т.п.). При оформлении подписки на получение ресурсов из каталога пользователь должен иметь возможность сформировать собственные фильтры для отбора интересующих его веб-ресурсов. Такой фильтр может быть задан как для рубрик (отбор ресурсов по тематическому и/или географическому

принципу), так и для атрибутов ресурсов (отбор ресурсов по доступности, популярности, приоритету, дате создания/обновления метаописания и т.д). В принципе, поддержка фильтрации контента является несложной с технической точки зрения задачей до тех пор, пока режим фильтрации и передачи данных имеет четко выраженный односторонний характер (от поставщика к потребителю).

Заметную часть подписчиков каталога составляют потребители, которые имеют собственные массивы метаданных. Как правило, для обслуживания этой категории пользователей одного механизма фильтрации метаданных оказывается недостаточно, и приходится поддерживать различные схемы интеграции между массивами метаданных каталога и подписчика. Возможен вариант, когда подписчик соглашается на полную интеграцию своего массива с метаданными каталога, предполагая в последующем отказаться от выполнения функции ведения собственного массива метаданных, и полностью перейти на обслуживание своих потребностей за счет массива метаданных каталога.

В другом случае, подписчик может запросить метаописания тех ресурсов, которые отсутствуют в его информационном массиве. Тогда подписчик предоставляет полный перечень адресов (URL) интернет-ресурсов, описания которых у него имеются (возможно, в режиме регулярного пополнения). Этот перечень используется системой отгрузки как дополнительный фильтр для отбора передаваемых клиенту метаданных (т.е. передаются только ресурсы, которые не вошли в перечень).

Еще один вариант интеграции предполагает, что подписчик хочет расширить атрибутный состав своего метаописания за счет атрибутов, поддерживаемых в каталоге. В этом случае подписчик также должен предоставить полный перечень адресов (URL) интернет-ресурсов, описания которых у него имеются, а в ответ получает выгрузку, содержащую только те атрибуты ресурсов, которые он заказал, причем в выгрузку попадают только ресурсы, содержащиеся в представленном перечне.

С технической точки зрения, наиболее сложной представляется схема, при которой процесс обмена данными носит итеративный характер, и который управляется персоналом (операторами) заказчика. Для обеспечения подобных схем взаимодействия приходится реализовывать специализированные утилиты, выполняющие следующие операции:

- **Первоначальная отгрузка ресурсов** – формирует выгрузку ресурсов, применяя специфицированный подписчиком фильтр и, дополнительно, выбирая только те ресурсы, которые не вошли в предыдущие выгрузки.
- **Прием одобренных и отвергнутых описаний ресурсов –** принимает от подписчика пакет, в котором содержатся все ресурсы, которые были в соответствующем пакете первоначальной выгрузки, где каждый ресурс имеет статус «принят», «на доработку» или «отвергнут». Статус «на доработку» означает, что заказчика не совсем удовлетворяет описание ресурса. В этом случае, дополнительно передается атрибут с текстом

замечания. Предполагается, что редактор каталога устранит высказанные замечания, после чего описание ресурса будет повторно передаваться заказчику (в этом месте вполне возможно возникновение циклов).

- **Отгрузка доработанных ресурсов** – формирует выгрузку ресурсов, описания которых были доработаны в соответствии с пожеланиями заказчика.

2.3 Режимы и механизмы распространения информации

Существует два принципиально различных способа, при помощи которых содержание каталога (полностью или частично) может быть предоставлено в распоряжение подписчика: один из них предполагает непосредственную передачу контента (отгрузка данных), а другой – предоставление доступа к сервисам.

Для передачи данных могут быть задействованы следующие механизмы:

SQL-интерфейс к удаленной базе данных каталога. Данное решение позволяет подписчику работать с (реструктурированными и отфильтрованными) исходными данными на уровне SQL запросов. Такого рода низкоуровневый доступ дает, с одной стороны, существенную гибкость подписчику, с другой стороны, требует некоторых затрат на программирование и поддержание системы с его стороны. Это решение также предполагает наличие высокоскоростного канала связи между сайтами подписчиков и базой данных каталога (например, это достигается путем организации хостинга в рамках одного провайдера).

Синхронизация базы данных подписчика. Решение, которое может быть реализовано с применением стандартного механизма репликации данных. Обладает тем достоинством по отношению к предыдущему варианту, что временная потеря связи с базой данных каталога не отражается на работоспособности сервера подписчика – данные хранятся локально и периодически обновляются/ синхронизируются. Актуальность данных зависит от требований подписчика и достигается путем соответствующих настроек синхронизации. Важной особенностью данного способа является то, что передаётся только обновлённая часть данных.

Выгрузка данных в XML-файлы. Этот механизм является наиболее эффективным по соотношению цена/качество способом распространения каталога. С требуемой периодичностью (например, ежечасной, ежедневной, еженедельной) происходит выгрузка данных в том или ином виде либо на ftp/http сервер для последующего скачивания системой подписчика, либо перекачивается непосредственно на его сайт. Такая работа может выполняться в инкрементальном режиме, т.е. когда подписчику отгружаются только новые или измененные (с момента последней передачи) ресурсы, рубрики и связи между ними. Контент предоставляется в согласованных с подписчиком структурах, создаваемых на основе XML, который является универсальным форматом для передачи

данных и позволяет легко обмениваться информацией между разными приложениями.

Выгрузка данных в виде сгенерированного Веб-сайта . В этом случае заказчик получает уже готовый набор страниц в стандартных форматах – XML+XSL, DHTML, HTML+CSS, HTML. Выбор формата определяется заказчиком, страницы готовы к скачиванию и непосредственному встраиванию в сайт заказчика.

Очень перспективным способом публикации каталога, который способен привлечь массового подписчика, является переход от распространения каталога путем передачи данных к организации сервисов каталога, к которым подписчик получает доступ. В пользу этого варианта говорит то, что такое решение практически не требует затрат на разработку и/или интеграцию со стороны подписчика. При этом обеспечивается:

- полная интеграция в существующий дизайн и структуру сайта подписчика;
- расширение спектра информационного наполнения страниц сайта;
- привлечение потенциальных посетителей сайта;
- позиционирование сайта подписчика как мини-каталог с определенными тематическими ресурсами;
- минимизацию расходов на создание и поддержку собственного каталога.

К числу сервисов, при помощи которых подписчик публикует на страницах своего Веб-сайта информацию из каталога, можно отнести сервис импорта информационных блоков, в которых размещается отфильтрованные фрагменты существующей базы данных каталога. Информационные блоки могут быть следующих видов:

- **Отдельная страница.** Информация из базы данных каталога размещается на отдельных страницах, с сохранением дизайна сайта подписчика. На такой странице сохраняется базовая функциональность каталога: перемещение по рубрикам (при переходе по рубрикам фрагмента каталога пользователь остается на страницах веб-сайта подписчика); пролистывание порционного списка ресурсов больших рубрик; сортировку по приоритету, дате создания описания ресурса, названию и т.п.
- **Блок-информер.** Информация из базы данных каталога размещается в виде отдельных встроенных блоков на существующих страницах сайта подписчика. Информер представляет собой усеченную версию фрагмента каталога, в котором поддерживается два способа отображения контента. А именно, либо в виде динамического списка, состоящего из нескольких ссылок, появляющихся в результате случайной выборки из тематического подмножества, выбранного подписчиком; либо в виде порционного списка ресурсов с возможностью его листания. Привлекательность данного вида блока

состоит в отсутствие необходимости встраивания в структуру веб-сайта новых страниц. При этом допускается самостоятельное регулирование графических размеров блока.

- **Бегущая строка.** Информация из базы данных отображается в виде текста бегущей строки, которая может быть размещена в любой части страницы сайта подписчика. Такая строка формируется на основании заданного подписчиком фильтра (где, в частности, может быть задан порядок сортровки) и представляется как зацикленный список описаний ресурсов.

Содержимое блоков отгружается на сайты подписчиков непосредственно из каталога, однако явно это не видно конечному пользователю. Механизм работы данного сервиса аналогичен баннерным сетям, когда на страницах размещается скриптовый файл, предоставляемый каталогом. Важным преимуществом для подписчика является также то, что он избавляется от необходимости хранить что-либо на своем сайте. Это решение должно оказаться полезным для тех, кто размещает свои сайты на виртуальных серверах провайдеров.

Заключение

Создание развитой инфраструктуры для представления и обмена метаданными является одним из приоритетных направлений совершенствования современной глобальной сети. Наиболее известными шагами в этом направлении со стороны World Wide Web Consortium (W3C) стало появление стандарта представления метаданных RDF (Resource Description Framework [1]) и открытая в 2001 году инициатива "Semantic Web" [2], призванная скоординировать усилия по созданию прикладных RDF-ориентированных инструментов и технологий. В перспективе, в результате реализации этих инициатив, значительная часть ресурсов Интернет получит информативное структурированное описание в сети, созданное владельцами ресурсов. Это позволит существенно облегчить задачу автоматического формирования массивов метаданных в каталогах. С другой стороны, решение задач проверки достоверности, оценки качества ресурса и его позиционирования в соответствующем классе ресурсов, останется в ведении экспертов каталога.

В настоящее время многие информационные центры, занимающиеся сбором и распространением метаданных, проявляют активную заинтересованность в организации взаимодействия с целью обмена имеющимися у них фондами. Как правило, в основе такой интеграции фондов лежит выработка стандарта на формат для представления метаданных, одновременно с унификацией массивов нормативно-справочной информации.

Существует ряд предметных областей, где взаимодействие каталогов строится по такой схеме. В наибольшей степени в этом направлении сегодня

продвинулись библиографические каталоги, где широко используется организация обмена на основе стандартов UNIMARC (расширенного для представления описаний электронных документов в ISBD(ER) [3]) и Dublin Core [4]. Однако можно заметить, что такой подход применим не для всех типов ресурсов. С одной стороны, существуют такие "многоаспектные" типы ресурсов (например, понятие Веб-сайта или вычислительного ресурса сети), для которых вообще не ясна перспектива выработки единого способа описания. С другой стороны, появляются новые, быстро эволюционирующие типы ресурсов (например, мультимедиа-ресурсы, интерактивные сервисы сети и т.п.), разработка стандартов для которых, в силу их динамической природы и новизны, не успевает за темпами развития данных предметных областей.

Предложенная в настоящей работе архитектура ориентирована на то, чтобы позволить заинтересованным организациям решить задачу создания каталогов метаданных, способных адекватным образом описывать такие многоаспектные и динамичные категории ресурсов, какими являются ресурсы Интернет и, с другой стороны, за счет привлечения широкого круга подписчиков, превратить такой каталог в экономически эффективное предприятие. Несомненно, что удовлетворение функциональных требований, возникающих при многоцелевом режиме эксплуатации каталога является сложной с технической точки зрения задачей, особенно в свете необходимости поддержки активной эволюции структур данных в условиях непрерывно идущего редакционного процесса. При этом требуется обеспечить безконфликтную параллельную работу процессов модификации схемы, ведения данных, а также обмена данными между каталогом и узлами-подписчиками.

Литература

[1] Resource Description Framework (RDF), http://www.w3.org/RDF
[2] W3C Semantic Web Activity, http://www.w3.org/2001/sw/
[3] International Standard Bibliographic Description for Electronic Resources, 1997, http://www.ifla.org/VII/s13/pubs/isbd.htm
[4] Dublin Core Metadata Initiative (DCMI), http://dublincore.org