ДОКУМЕНТАЛЬНЫЕ ИСТОЧНИКИ ИНФОРМАЦИИ

УДК 004.738.5

О. В. Барышева

Интернет — Метаданные — Dublin Core*

Рассматриваются принципы описания электронных документов, получившие название "Даблин Кор" (Дублинское ядро), разрабатываемые для международного использования Автоматизированным библиотечным центром с интерактивным доступом (г. Дублин, шт. Огайо, США, OCLC — Online Computer Library Center).

Вопросы поиска и нахождения нужной информации в среде традиционных документов к настоящему моменту можно признать (с некоторой натяжкой) решенными удовлетворительно. Во всяком случае, читатель, приходящий в библиотеку, использует справочный аппарат или прибегает к помощи библиографов, но, так или иначе, получает искомый документ (в крайнем случае, сведения об его местонахождении).

ОТ БИБЛИОГРАФИЧЕСКИХ ОПИСАНИЙ К МЕТАДАННЫМ

Электронные каталоги или базы данных на отдельные части фондов, которыми уже обзавелись крупнейшие библиотеки нашей страны, позволяют получать более точные и полные результаты библиографического поиска. Полнотекстовые базы данных как более быстрые и сложные поисковые системы обычно снабжаются изрядным количеством служебной информации, позволяющей идентифицировать документ с запросом. Кроме того, обычно база наполняется однотипными документами (по тематике, виду, другим содержательным или формальным признакам). Эти однотипные документы представлены в базе в виде упорядоченного и достаточно жестко структурированного массива данных. Само тело документа (например, текст статьи, звуковой ряд музыкального произведения etc.) крайне редко выступает в качестве поискового поля.

Мы формируем запрос на основе произвольного набора имеющихся у нас данных, а поиск сводится к проверке наличия соответствующих данных о документах в соответствующих (аналогичных поисковым) полях. Таким образом, даже полнотекстовые базы данных состоят как минимум из двух частей: описания данных (с жесткой структурой и фиксированными полями) и собственно данных (т. е. полных текстов). Электронные же каталоги представляют собой лишь первую составляющую, роль второй выполняют собственно фонды библиотек, хранилища собственно данных. Если для связи этих двух частей уже давно и традиционно принято пользоваться системой шифров, то для описания документов в электронных каталогах был разработан специальный формат — структурированный поблочно набор полей для машиночитаемой каталогизации (MARC — Machine-readable cataloguing).

ПОЛНОТЕКСТОВАЯ БАЗА ДАННЫХ — ПОЛНОТЕКСТОВЫЙ ПОИСК В БИБЛИОТЕКЕ

Уже обработанные (имеющие вид базы данных со своим поисковым интерфейсом) материалы чаще всего выпускаются на CD-ROM за отдельные промежутки времени. Отдельные актуально обновленные и абсолютно полные БД можно найти и в Интернет, также как и электронные каталоги. Но помимо этого Интернет наполнена массой разрозненных материалов, не поддающихся структурированию, так как они могут иметь, например, разных владельцев, разную языковую природу, неодинаковое программное обеспечение, требуемое для их "чтения". Что делается в этом случае?

Мы не будем пытаться в очередной раз создавать классификацию поисковых машин или роботов, работающих в сети. Просто посмотрим, где хранятся собственно данные (в нашем случае — первичные документы), где — их описания (которые служат базой для поиска) и как они связываются между собой. Для удобства понимания и единства терминологии с общепринятой (но до сих пор не стандартизованной) мы будем использовать следующие понятия:

данные — для обозначения полных текстов или первичных документов;

метаданные — "данные о данных" для обозначения описаний.

*Dublin Core Metadata for Resource Discovery. Internet RFC 2413 <http://www.ietf.org/rfc/rfc2413.txt>

Возвращаясь к поставленному выше вопросу о том, как искать разнородные документы в глобальной сети, отметим, что для этого точно так же используются описания или метаданные. Но создаются они по-разному. Все зависит от того, в какой поисковой системе создатель или владелец данных хочет их представить.

В системы с ручной индексацией необходимо, что явствует из их названия, вручную вводить метаданные по шаблону в поля, перечень которых предлагается системой. Данные же при этом хранятся вне этой системы, т. е. на сервере владельца или создателя.

Системы с автоматической индексацией предполагают наличие метаданных в теле документа, т. е. практически данные и метаданные выглядят как единый документ. Разница состоит лишь в том, что поля метаданных невидимы для потребителя данных (читателя). Для примера приведем метаданные этой статьи (в разметке HTML — HyperText Markup Language)*:

<!DOCTYPE HTML PUBLIC "—//W3C//DTD HTML 3.2 Final//EN">
<HTML>
<HEAD>
*<TITLE>***Интернет-Метаданные-Doublin Core***</TITLE>*
<META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=windows-1251">
<META NAME="GENERATOR" CONTENT="Mozilla/4.04 [en] (Win95; I) [Netscape]">
<META NAME="Generator" CONTENT="Winword 97">
*<META NAME="Author" lang="ru" CONTENT="***Ольга Барышева***">*
*<META NAME="Keywords" lang="en" CONTENT="***documents, descriptions, metadata***">*
*<META NAME="Description" lang="ru" CONTENT="***Эта статья посвящена формату метаданных для описания электронных ресурсов***">*
*<META NAME="Description" lang="en" CONTENT="***It is an article about the Dublin Core metadata format***">*
</HEAD>

(*Жирным выделены элементы, создаваемые вручную, остальное — тэги HTML — может быть создано средствами HTML-редактора*).

Далее мы выпускаем документ, снабженный приведенными выше метаданными, в глобальную сеть. Роботы поисковых машин, ориентируясь на имена тэгов HTML, извлекают из него метаданные в свою базу данных, по которой осуществляется поиск. Недостатки такой схемы работы очевидны: разные люди по-разному интерпретируют значения мета-тэгов (*<META NAME>*). Кроме того, приведенный выше набор имен (*Content-Type, Generator Content, Generator, Author, Keywords, Description*) не является конечным и, к сожалению, не является обязательным. Не разработано и строгих правил приведения данных (например, порядок имени и фамилии автора). Мы умолчим также о возможных случайных и сознательных искажениях содержания документов в метаданных (а пользователи не имеют возможности даже с целью верификации получить сведения о создателях последних), других способах достижения "максимальной популярности" своих данных, так как это не относится к рассматриваемой теме. По всем приведенным выше и ряду других причин поиск в Интернет в большинстве случаев не приносит ожидаемых результатов.

Имея перед собой цель исправить существующее положение, в марте 1995 г. в г. Дублине (штат Огайо) 52 ученых и специалиста в области библиотечного дела, информатики и смежных дисциплин приняли модель описания электронных ресурсов, которая получила название Дублинского ядра элементов метаданных (Dublin Metadata Core Elements)**. Эволюция документа, описывающего эту модель, а точнее, уже формат, происходила и происходит под эгидой Dublin Core Metadata Initiative (далее DCMI)***. В 1999 г. DCMI рекомендовала набор из 15 элементов: название, создатель, предмет, описание, издатель, соисполнитель, дата, тип, формат, идентификатор, источник, язык, отношение, охват, права (Dublin Core Metadata Element Set, далее DCES в версии 1.1). Сведения об этих элементах приведены в табл. 1.

Мы не будем подробно останавливаться на том, что словосочетанием "Dublin Core" (далее DC) называют как DCMI в целом, так и отдельные рабочие группы, т. е. любую *деятельность*, связанную с разработкой и внедрением DCES. Чаще всего (и мы будем придерживаться этой практики) словосочетание "Dublin Core" обозначает *формат* для описания электронных документов.

*HTML — HyperText Markup Language Home Page <http://www.w3.org/MarkUp/>.

**DCES — Dublin Core Metadata Element Set, Version 1.1: Reference Description <http://purl.org/DC/documents/rec-dces-19990702.htm>.

***DCMI — Dublin Core Metadata Initiative Home <http://purl.org/DC/>;

Таблица 1

Элементы Dublin Core

Имя элемента	Идентификатор	Дефиниция (определение)	Комментарии
Название	Title	Имя, данное ресурсу	Обычно *название* — это имя, под которым ресурс известен.
Создатель	Creator	Лицо (лица), несущее первичную ответственность за создание и содержание ресурса	Примеры *создателя* включают персону, организацию или службу. Обычно имя создателя следует использовать для индикации объекта описания.
Предмет и ключевые слова	Subject	Предметная область, определяющая содержание ресурса	Обычно *предмет* выражается с помощью *ключевых слов*, ключевых фраз или кодов классификаций, которые описывают тематическую принадлежность ресурса.
Описание	Description	Сообщение о содержании ресурса	*Описание* может включать (но не ограничивается): реферат, оглавление, ссылки на графическое представление содержания или простое текстовое изложение содержания.
Издатель	Publisher	Лицо (лица), несущее ответственность за ввод ресурса в обращение	Примеры *издателя* включают персону, организацию или службу. Обычно имя издателя следует использовать для индикации объекта описания.
Соисполнитель	Contributor	Лицо (лица), несущее ответственность за содействие в создании содержания ресурса	Примеры *соисполнителя* включают персону, организацию или службу. Обычно имя соисполнителя следует использовать для индикации объекта описания.
Дата	Date	Дата, связанная с событием в жизненном цикле ресурса	Обычно дата ассоциируется с созданием или доступностью ресурса. Рекомендуемое для практического использования при кодировке значение *даты* определено в профиле ИСО 8601 и поддерживает формат ГГГГ-ММ-ЧЧ.
Тип ресурса	Type	Свойство или жанр содержания ресурса	*Тип* включает термины, включающие общие категории, функции, жанры или объединенные уровни содержания. Для практического использования рекомендуется выбирать значение из контролируемого словаря (e.g. DCT*). Для описания физического или цифрового представления ресурса используется элемент *формат*.
Формат	Format	Физическое или цифровое представление ресурса	Обычно *формат* может включать тип копии (медиа-тип) или величину ресурса. *Формат* может использоваться для определения технического и программного обеспечения или другого оборудования, необходимого для отображения или управления ресурсом. Примеры величины включают размер и продолжительность. Для практического использования рекомендуется выбирать значение из контролируемого словаря (e.g. MIME*).
Идентификатор ресурса	Identifier	Однозначная ссылка на ресурс в пределах данного контекста	Для практического использования рекомендуется идентифицировать ресурс посредством строки или числа, соответствующего формальной идентификационной системе (URI*, URL*, DOI*, ISBN*).
Источник	Source	Ссылка на тот ресурс, из которого извлечен настоящий	Настоящий ресурс может быть извлечен из *источника* целиком или частично. Для практического использования рекомендуется идентифицировать ресурс посредством строки или числа, соответствующего формальной идентификационной системе.
Язык	Language	Язык интеллектуального содержания ресурса	Для практического использования рекомендуется значение элемента *язык*, определяемое RFC 1766, включающим двухбуквенные коды языков (взятые из стандарта ИСО 639), со следующими факультативно двухбуквенными кодами стран (взятыми из стандарта ИСО 3166*). Например, "en" — для английского, "fr" — для французского,"en-uk" — для английского, используемого в Великобритании.
Отношение	Relation	Ссылка на родственные ресурсы	Для практического использования рекомендуется идентифицировать ресурс посредством строки или числа, соответствующего формальной идентификационной системе.

Имя элемента	Идентификатор	Дефиниция (определение)	Комментарии
Охват	Coverage	Протяженность и границы содержания ресурса	*Охват* обыкновенно включает пространственное местонахождение (название местности или географические координаты), временной промежуток (временная метка, дата или диапазон дат) или юрисдикцию (такую, как названное административное подразделение). На практике рекомендуется выбирать значение из контролируемого словаря (например, Тезауруса географических названий), т. е. целесообразнее использовать названия местностей и периодов времени вместо цифровых идентификаторов (таких, как системы координат или диапазоны дат)
Правовое регулирование	Rights	Информация о правах по ограничению доступа и охране ресурса	Обычно элемент *права* содержит положение о правовых нормах, регулирующих функционирование ресурса, или ссылку на службу, предоставляющую эту информацию. Правовая информация обычно включает сведения о правах на интеллектуальную собственность, авторском праве и других имущественных правах. Отсутствие элемента *права* не может являться основанием для каких-либо предположений о правовом статусе относительно ресурса

*DCT — List of Resource Types: Dublin Core Draft Working Group Report <http://purl.org/DC/documents/wd-typelist.htm>; MIME — Internet Media Types <http://www.isi.edu/in-notes/iana/assignments/media-types/media-types>; URI — Uniform Resource Identifiers: Generic Syntax, Internet Draft Standard http://www.ics.uci.edu/pub/ietf/uri/rfc2396.txt>; URI, URL-Naming and Addressing: URIs, URLs, ... <http://www.w3.org/Addressing/>; URL — Uniform Resource Locator Specification <http://www.w3.org//Addressing/URL/Overview.html>; DOI — The Digital Object Identifier <http://www.doi.org/>; ISBN — International Standard Book Numbering <http://www.reedref.com/standards/>; ISO 3166 2-letter country codes <http://www.w3.org/International/O-misc-iso3166.html>;

ЧТО ТАКОЕ DUBLIN CORE?

DC — это набор метаданных, использование которых должно облегчать поиск электронных ресурсов (прежде всего, сетевых). Наибольшую заинтересованность в развитии DC пока что проявляют библиотеки, научные и культурные организации, правительственные агентства, а также коммерческие фирмы.

Целью развития DC является создание такого механизма, который при минимальных затратах на описание позволял бы искать и находить данные вне зависимости от языковой принадлежности, тематики и места их нахождения. Для этого DCMI должны быть решены и решаются следующие задачи: создание национальных и полиязыковых баз данных (регистраторов ресурсов) на основе DC, создание шаблонов для описания, создание руководств и содействие распространению формата в сетевом сообществе, разработка идентифицирующих меток для верификации метаданных. При решении этих и многих других задач DCMI руководствуется следующими принципами.

Принцип *простоты*. DC рассчитан как на специалистов по описанию ресурсов, так и на не-каталогизаторов. Ориентиром должна быть интуитивная понятность описания для любого пользователя вне зависимости от уровня и профиля его образования. В противном случае мы опять получим систему, работа с которой будет возможна лишь при помощи посредников. Кроме того, невозможным будет соблюдение следующего принципа — принципа *возможности семантического взаимодействия*. Метаданные DCES должны являть собой понятный для людей и машиночитаемый набор дескрипторов, с помощью которого можно описывать данные из любой тематической области или науки. Принцип *международного согласия* позволяет представителям различных стран и континентов не только принимать участие в работе DCMI, но создавать национальные и международные системы на базе DC, которые могут успешно взаимодействовать. Принцип *расширяемости* не только подтверждает факт, что DC является более гибким, по сравнению с MARC-форматом, средством описания. Кроме того, с помощью *DC* могут быть описаны данные более сложные и разнообразные, чем поток традиционных документов. Последним и наиболее важным является принцип возможности предоставления метаданных во World Wide Web. Соблюдением этого принципа достигается общедоступность метаданных и поиска по ним, доступ к условиям получения с использования самих данных.

В настоящее время W3 Консорциум* начал работы по внедрению структур метаданных в архитектуру Всемирной паутины. Система описания ресурсов (Research Description Framework — RDF**) спроектирована таким образом, чтобы удовлетворять разнообразные потребности в сведениях о данных, варьирующиеся в зависимости от

*W3C — World Wide Web Consortium <http://www.w3.org/>;
**RDF — Resource Description Framework <http://www.w3.org/RDF>;

поставщиков / провайдеров информации, а также учитывать перспективы развития электронных библиотек как составной части инфраструктуры Всемирной паутины.

СТРУКТУРА И ИНТЕРПРЕТАЦИЯ DC

Формат DC — перечень полей и правил их заполнения или (с лингвистической точки зрения) набор элементов и правил построения высказываний. С его помощью (на его основе) создается описание в формате DC, которое, в идеале, производится самими создателями документов.

Иначе говоря, DC являет собой язык для специальных целей (создания метаданных), который должен быть доступен для понимания человеком и машинной интерпретации.

Для поиска и функционирования в компьютерной сети DC-метаданные вместе с документами (данными) или отдельно от них должны быть преобразованы / переведены на язык, доступный для анализа / понятный для читающего Интернет-документы устройства — браузера. Таким образом, DC-формат определяет семантические области, которые надлежит отразить в описаниях (абстрактную семантику), DC-описание есть набор конкретно-семантических элементов, а Интернет-проекция этого описания — тот же набор, но уже синтаксически (с помощью языка разметки, например, HTML, SGML* и др.) оформленный. Например:

* наличие элемента *Создатель* говорит о том, что в описание по формату DC должны быть внесены сведения о человеке, коллективе или организации, которые явились авторами описываемого документа или объекта (иначе говоря, метаданные DC должны включать сведения о создателе данных);
* полученные сведения должны выглядеть в DC-описании как сумма идентификатора данного элемента и его значения — *DC.Creator=Перов А.*;
* для поиска в Интернет эта запись будет выглядеть как сумма DC-описания и тэгов HTML *<META NAME="DC.Creator" CONTENT="Перов А.">* **.

Основу формата составляют описанные 15 **элементов**. Каждый из них имеет свое название и текстовый идентификатор для однозначного смыслового соотношения. Например:

* элемент *ПРЕДМЕТ* имеет название *ПРЕДМЕТ И КЛЮЧЕВЫЕ СЛОВА* и идентификатор *Subject*.

Кроме того, многие элементы имеют дополнительные квалификаторы. Эти, так называемые **квалификаторы элементов**, служат для семантической очистки последних. Практически их можно интерпретировать как видовые по отношению к элементам, т. е. квалификатор всегда у́же элемента. Например, *Иллюстратор* — есть тип *Создателя*. Каждый квалификатор элемента должен иметь свое имя и четкое определение. Например, один из квалификаторов элемента *Охват*:

Имя (и идентификатор): periodName
Определение: Название периода во времени.
Комментарий: Названия Периодов должны всегда, если это возможно, выбираться из контролируемых словарей с указанием названия контролируемого словаря, из которого составителем описания был извлечен квалификатор значения.

При использовании квалификаторов должен выполняться следующий принцип (dumb down principle): если клиент не распознает квалификатор элемента, он может его просто проигнорировать, но при этом высказывание остается истинным.

Следующей составной частью DC являются **квалификаторы значений**, назначающие схемы кодировки или контролируемые словари, которые помогают определить значение конкретного элемента и одновременно облегчают анализ метаданных. Например,

Схемы кодировки отображают правила приведения и интерпретации данных. Стандартным примером, наиболее ярко иллюстрирующим необходимость наличия квалификаторов элемента и значения, является приведение значения Даты. Например, запись без квалификаторов:

DC.Date=19991206 может быть прочитана так: *"12 июня (либо 6 декабря) 1999 года является датой, связанной с функционированием (созданием, доступом, обновлением, выходом в свет и т. д.) описываемого документа"*; тогда как запись с квалификаторами элемента и значения (схемой)

DC.Date.Created Scheme=ISO8601 Content= =19991206 однозначно читается как: *"описываемый документ был создан 6 декабря 1999 года"*.

Контролируемые словари выполняют ту же функцию, что и схемы кодировки и имеют общий идентификатор — Scheme. Например, указание на библиотечную классификацию *Scheme=UDC* или *Scheme=DDC* позволяет верно прочитать систематический индекс для определения тематики документа, как и ссылки на тезаурусы, которые в качестве контролируемых словарей позволяют адекватно переводить предметные рубрики с одного языка на другой, однозначно интерпретируя их семантику, что значительно улучшает возможность взаимодействия при многоязычном поиске.

Каждый квалификатор значения должен иметь уникальное название и ссылку на подробную информацию о той или иной схеме кодировки или контролируемом словаре, так как они, безусловно, должны быть открыты для доступа.

СОЗДАНИЕ И ОФОРМЛЕНИЕ DC-ОПИСАНИЙ

Для того чтобы упростить задачу создания описаний, используются специальные прикладные программы, включающие пользовательские шаблоны для ввода описаний. Это позволяет создавать корректные с лингвистической точки зрения метаданные на основе DC без знания используемого языка разметки. В настоящее время наиболее популярным во всем мире продолжает оставаться HTML-кодировка в версиях HTML 3.2 и 4.0, тогда как наиболее перспективным является использование XML*** (eXtensible Markup Language) и RDF (Research Description Framework).

*SGML — The SGML/XML Web Page <http://www.oasis-open.org/cover/sgml-xml.html>.
**Encoding Dublin Core Metadata in HTML <http://www.ietf.org/rfc/rfc2731.txt>.
*** XML — Extensible Markup Language <http://www.w3.org/XML/>.

7 DATE (current date=Datum)
Der Eintrag des Datums wird automatisch aus dem Maschinendatum erzeugt.

8 TYPE (Art des Dokuments):
Dissertation

9 FORMAT (Datentechnisches Format des Dokuments, MIME type):
text/html MIME Type

10 IDENTIFIER (ISBN, ISSN, URL o.?. des **vorliegenden Dokuments** betr. eindeutiger Identifikation):
URL URN
ISBN ISSN

11 SOURCE (SWB-ID-Nr der Titelaufnahme des **vorliegenden Dokuments** in der SWB-Verbunddatenbank):
SWB-ID-Nr (8-stellig)

11 SOURCE (Werk, gedruckt oder elektronisch, **aus dem das vorliegende Dokument stammt**):
Verbale Form des Werks
ISBN oder ISSN

12 LANGUAGE (Sprache des Inhalts des Dokuments)
deutsch englisch franz?sisch spanisch italienisch
Liste mit anderen Sprachen

Выше приведен фрагмент пользовательского шаблона, размещенного на сервере Центра библиотечных услуг Баден-Вюртемберга*.

Хранятся DC-описания отдельно в виде базы данных. При поиске выдается адрес хранения самого документа и переход к нему возможен напрямую по указанной ссылке. Таким образом, создается Web-ориентированная база данных на распределенные ресурсы с единым механизмом поиска вне зависимости от содержания и типа документов, языковой принадлежности и кодировки, что значительно упрощает механизм поиска.

Ниже приводятся результаты обработки шаблона, в который введены сведения о данной статье (полужирным выделены элементы DC, курсивом – значения, которые вводятся вручную в шаблон).

```
<!DOCTYPE HTML PUBLIC "-//IETF//DTD HTML 4.0//EN">
<HTML>
<HEAD>
<META NAME="DC.Title" CONTENT="Интернет-Метаданные-Dublin Core">
<META NAME="DC.Title.Alternative" CONTENT="Internet-Metadata-Dublin Core">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#title">
<META NAME="DC.Creator.PersonalName"CONTENT="Barysheva Olga">
<META NAME="DC.Creator.PersonalName.Address" CONTENT="edd@nlr.ru">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#creator">
<META NAME="DC.Subject" CONTENT="Internet">
<META NAME="DC.Subject" CONTENT="resource description">
<META NAME="DC.Subject" SCHEME="LCSH" CONTENT="Information science">
<META NAME="DC.Subject" SCHEME="TGN" CONTENT="Russia">
<META NAME="DC.Subject" SCHEME="UDC" CONTENT="681.3">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#subject">
<META NAME="DC.Description" CONTENT="Статья посвящена формату метаданных для описания электронных ресурсов.">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#description">
<META NAME="DC.Publisher" CONTENT="ВИНИТИ">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#publisher">
<META NAME="DC.Contributor" CONTENT="DC7">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#contributor">
<META NAME="DC.Date.Created" SCHEME="ISO8601" CONTENT="1999-11-11">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date">
<META NAME="DC.Type" CONTENT="Text.Article">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#type">
<META NAME="DC.Format" SCHEME="IMT" CONTENT="application/msword">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#format">
<META NAME="DC.Identifier" CONTENT="http://www.viniti.ru/nti/articles/barysh.doc">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#identifier">
<META NAME="DC.Source" CONTENT="Научно-техническая информация">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#source">
<META NAME="DC.Language" SCHEME="ISO639-1" CONTENT="ru">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#language">
```

* DC-Template: DC-Meta-Maker des Bibliotheksservice-Zentrum Baden-Wuerttemberg <http://www.bszbw.de/diglib/medserv/konvent/metadat/dcmakelt.html>.

```
<META NAME="DC.Relation" SCHEME="URL" CONTENT="http://purl.org/DC/documents/rec-dces-19990702.htm">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#relation">
<META NAME="DC.Coverage.PlaceName" CONTENT="Russia">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#coverage">
<META NAME="DC.Rights" CONTENT="Доступен без ограничения">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#rights">
<META NAME="DC.Date.X-MetadataLastModified" SCHEME="ISO8601" CONTENT="1999-12-27">
<LINK REL=SCHEMA.dc HREF="http://purl.org/metadata/dublin_core_elements#date">
<HEAD>
```

В ряде случаев меню предлагается в неформализованном виде, т. е. в шаблонах используется лексика так называемого "естественного" языка, которая затем переводится на язык для специальных целей в соответствии с используемой схемой. Так, например, часто происходит выбор языка, что можно видеть, если сравнить шаблон с получаемой записью. Следующая за каждым элементом метаданных ссылка (*<LINK REL>*, в целях экономии места в примере повторы опущены) дает возможность проверить правильность употребления элемента и его квалификаторов, а также при необходимости получить дополнительную информацию.

ИСПОЛЬЗОВАНИЕ DC И ЕГО ПЕРСПЕКТИВЫ В РОССИИ

Грубо обобщив статистические данные из разных отечественных и зарубежных источников, можно сказать, что в России на момент наступления 2000 г. существует приблизительно 50 тыс. уникальных Web-серверов. Поиск по ним ведется с помощью максимум 15 поисковых систем и каталогов. Это если говорить о русском языке. В ином случае пользователь Интернет волен выбирать из более, чем 1500 проводников в глобальную компьютерную, а самое главное — информационную сеть.

Для более или менее успешного, а вернее, удовлетворительного поиска необходимо знание особенностей наполнения и поискового механизма (правил формулирования запроса, способов ранжирования документов при выводе и т. п.) поисковых систем. Беглый сравнительный анализ показывает, что внутренняя структура, применяемые операторы и поддерживаемые функции даже отечественных систем сильно разнятся. Рядовой пользователь должен прежде всего найти (что не всегда просто сделать), прочитать, понять (а объяснения не всегда понятны) и по возможность выучить (а объем иногда превышает разумные пределы) описание языка запросов той или иной конкретной системы. Только после этого можно решить, подходит ли она для решения тех задач, которые стоят перед пользователем.

Кроме того, хорошо бы до проведения первой сессии поиска иметь сведения о глубине индексирования и скорости поиска, а также о том, какой смысл создатели системы вкладывают в термин "релевантность". Можно еще долго перечислять недостатки систем и неудобства, возникающие при работе с ними. Безусловно, более привычен и интуитивно понятен поиск по электронному каталогу библиотеки. Обусловлено это, как было отмечено выше, наличием жесткой и обязательной структуры данных о документе.

Dublin Core был разработан для описания электронных документов в электронной среде аналогично тому, как MARC-формат был разработан для описания традиционных документов в электронной среде. DC не есть обязательный формат. Интернет по природе своей не может иметь ничего обязательного. Поэтому применение и самого формата, и поддерживающих его прикладных программ факультативно. Каждый член виртуального сообщества принимает решение сам — нужен ему поиск удовлетворительного качества или нет. На настоящий момент более 20 стран имеют переводы DC на свои языки, а некоторые (например, Финляндия) даже национальные версии.

В России решение об использовании DC на национальном уровне (для государственных организаций, специализирующихся на обработке информации, таких как центры НТИ и библиотеки) пока что не принято. Но работы по ознакомлению с DC стали появляться уже с 1998 г., а в 1999 г. представители России впервые получили возможность обменяться информацией с коллегами не по электронной почте, приняв участие в 7-м рабочем семинаре DCMI*, прошедшем в октябре во Франкфурте-на-Майне.

Представляется логичным дальнейшее ведение этой работы в России по следующим направлениям:

* перевод формата Dublin Core с полным набором квалификаторов, который появится 1 января 2000 г., на русский язык и его распространение;
* создание единых для библиотек и центров НТИ шаблонов с использованием HTML, XML, RDF, размещение их на серверах крупнейших организаций и обновление по мере появления новых версий DC и языков разметки;
* участие в дальнейших работах DCMI, по разработке и обновлению формата DC;
* проведение обучающих и информационных семинаров в различных регионах России.

ДОСТОИНСТВА И НЕДОСТАТКИ DC

Прежде всего, о достоинствах. Использование DC дает возможность быстрого, а главное, высококачественного поиска информации, поскольку имеет жестко структурированную *единую* схему создания и ввода описаний электронных документов. При необходимости можно расширить спектр вводимых реквизитов за счет как количества полей, так и введения дополнительных (для решения конкретных задач или специфических видов материалов) квалификаторов. Кроме того, отдельные базы данных на основе DC всегда могут быть объединены в одну не только механически, но возможно осуществление поиска по распределенной базе. Если та

*DC7 — The 7th Dublin Core Metadata Workshop <http://www.ddb.de/partner/dc7conference/results.htm>.

или иная организация, ведущая работу с электронными документами, хочет применять DC для описаний, но нуждается в сборе иных дополнительных сведений о документах, т. е. метаданных, она имеет возможность создать локальное расширение DC либо создать на его основе локальную базу данных для специальных целей. И в том, и в другом вариантах может быть сохранена возможность сетевого взаимодействия такой системы с остальными, поддерживающими "классический" DC (в версии 1.1). Это основные достоинства DC.

Но существуют еще и перспективные задачи, решив которые международное DC-сообщество получит новые преимущества. Прежде всего, это касается верификации описаний. Если пользователь работает со всем известными и общедоступными поисковыми системами типа Alta Vista, Hotbot, Yandex, Rambler и пр., у него нет никакой гарантии достоверности получаемой в результате поиска информации. Нет возможности выбора организаций, кроме посещения их Web-серверов, которые пользователь считает надежными поставщиками данных. Выбор группы доменов по географическому признаку, а в американской части Интернет — по типу организации эту задачу реально не решает. Каким же образом пользователь может найти в глобальной сети документ, который будет соответствовать его запросу не только по формально-содержательному критерию, но и по критерию истинности / достоверности описываемой информации? (Речь не идет о достоверности сведений, содержащихся в самом документе, но лишь об их надежности и соответствии метаданных собственно данным).

На последнем, 7-м рабочем совещании по Dublin Core предлагалось решить эту проблему с помощью Инициативы цифровой подписи*. Идея заключается в следующем: каждая организация будет иметь свою цифровую подпись, идентифицирующую составителя описаний. Тогда, проведя поиск, пользователь сможет выбирать из выданных документов (ориентируясь на идентифицируемые описания) те, которые обработаны в конкретных организациях, которым он, пользователь, доверяет. Возможно, подобные организации могут быть выбраны и на этапе формулирования запроса, до проведения сессии поиска.

Теперь о единственном, на наш взгляд, недостатке. Это касается элемента "Предмет и ключевые слова". Понятно, что в описаниях предпочтительно использовать нормативную лексику в соответствии с выбранным в качестве квалификатора значения контролируемым словарем. Для каталогизаторов тут проблем не возникает. А вот неспециалистам в области описания приходится трудно, так как они не владеют в совершенстве методикой предметизации и систематизации документов, а также не знают всего набора контролируемых словарей, используемого в современной библиотечной практике и в работе информационных центров и систем. Можно, конечно, ограничить значения этого элемента свободными ключевыми словами, но гораздо эффективнее был бы вариант предоставления через шаблон прямого доступа к наиболее популярным в конкретной области (например, Медицинским предметным рубрикам, Тезаурусу географических названий) или универсальным словарям (например, Предметным рубрикам Библиотеки Конгресса или Универсальной десятичной классификации), чтобы обработчик или создатель документа мог выбрать из соответствующего словаря соответствующее значение элемента.

ЗАКЛЮЧЕНИЕ

Традиционное библиографическое описание появилось, когда из-за большого количества традиционных документов осложнился процесс их разыскания, т. е. возникла необходимость регистрации и систематизации документопотока. Среда его обитания была бумажной, потому и сами документы и метаданные (каталожные описания) были бумажными.

В конце 60-х — начале 70-х гг. возник феномен единого электронного пространства, и среда обитания документов стала распределенной — электронно-бумажной. Традиционные бумажные документы стало возможным найти и в компьютерной сети в виде электронных копий, а не только на полках библиотек. Чтобы как-то разобраться в этом новом документопотоке был разработан MARC (для регистрации и поиска документов на бумажных носителях и их электронных копий).

Но появились и собственно электронные документы, которые не имеют бумажных аналогов, но могут быть связаны разными типами отношений с традиционными документами, а могут быть и абсолютно уникальными по своей природе. Для операций с ними было предложено использовать метаданные, которые получили окончательное оформление в виде Dublin Core, объединяющем в себе формат метаданных (описания электронных документов) и правила его составления. Можно с уверенностью сказать, что, несмотря на определенные недостатки, на сегодняшний день DC представляется наиболее разработанной и перспективной концепцией, описывающей методы обработки электронных документов и реально применимой практически для любых целей и для любых типов документов.

Материал поступил в редакцию 27.12.99.

*DSig — Digital Signature Initiative 1.0 <http://www.w3.org/DSig/>.