А. А. Воробьев, Ю. Г. Селиванова

МЕСТО И РОЛЬ БИБЛИОТЕК В SEMANTIC WEB: ПОДХОДЫ И РЕШЕНИЯ

Аннотация: *В статье рассматриваются основные принципы построения модели библиографических данных FRBR и концепция Semantic Web (семантической паутины), анализ которой позволил обосновать схожесть подходов в вопросах описания ресурсов, а также приводится информация о зарубежном опыте реализации представления библиографических данных инструментами Semantic Web.*

Ключевые слова: *Semantic Web, модель FRBR, RDF, Британская национальная библиография, авторитетный файл, Библиотека Конгресса США, модель библиографических данных, семантическая паутина, Интернет, библиотечное дело, открытые данные, связанные данные.*

Социальная миссия библиотек заключается в сохранении культурного наследия, представленного в виде книг, газет, журналов, картографических изданий, аудиовизуальных материалов, электронных ресурсов, и организации широкого доступа к нему. Как отмечают специалисты[1], современная библиотека расширила границы своей деятельности в условиях перехода из реального пространства в виртуальное. Наряду с созданием собственных электронных ресурсов, доступных в Web-среде, библиотеки предлагают различные сервисы по работе с ресурсами, принадлежащими другим субъектам информационного пространства.

[1] Концептуальная модель современной библиотеки: социально-философский анализ : автореф. дис. ... канд. филос. наук. Архангельск, 2007. 18 с. URL: http://tikunova-i.narod.ru/ni/koncept_avt.htm.

Для представления информации о ресурсах библиотечное сообщество использует комплекс лингвистических средств, позволяющих обеспечивать единообразное описание, индексирование библиотечных ресурсов в электронных каталогах (ЭК) и базах данных (БД) библиотек. Этот комплекс составляют: унифицированные форматы записи метаданных, правила и методики их формирования, контролируемые словари имен и наименований, предметных терминов.

Программа «Основные направления развития библиотечно-информационной сети ЛИБНЕТ на 2011–2020 гг.» предусматривает поэтапное обеспечение к 2020 г. свободного доступа всех граждан России к цифровой форме любого документа, хранящегося в любой библиотеке страны, через Интернет с соблюдением требований авторских прав. Планируется создание единой общероссийской библиотечной сети в Web-пространстве. К 2015 г. Россия должна иметь ЭК в 100% библиотек, и 50% документов должны быть переведены в цифровую форму. А к 2020 г. более 75% документов должны быть поэтапно оцифрованы и представлены пользователям в открытый доступ[1].

Размещение ЭК и цифровых ресурсов в сети Интернет стало общепринятой практикой. Но это отдельные разрозненные БД, не связанные между собой и БД других сообществ, сквозной поиск информации по которым затруднен из-за отсутствия единой точки доступа. Кроме того, библиотечные системы не позволяют напрямую интегрировать данные из внешних источников.

В связи с этим выполнение поставленной перед российскими библиотеками задачи потребует решения целого ряда вопросов, в том числе определения средств и механизмов представления данных в Web-среде с целью максимальной доступности для пользователей. Уже сейчас требуется не просто размещение данных, но и их проектирование в Web-среде.

В 1990-х гг. библиотечное и интернет-сообщества параллельно начинают разработку новых принципов описания ресурсов для Web. Необходимость таких разработок в библиотечной сфере была вызвана тем, что пользователи перестали обращаться к ЭК и фондам

---

[1] Основные направления развития Общероссийской информационно-библиотечной компьютерной сети ЛИБНЕТ на 2011–2020 гг. URL: http://www.nilc.ru/nilc/libnet-2011-2020.pdf.

библиотек из-за появления в сети Интернет огромного количества ресурсов, в том числе научных и образовательных. Пользователи нашли альтернативные варианты получения необходимой информации более быстрым и простым способом. Именно это подвигло библиотеки к поиску новых методов организации своей информации для интеграции ее в Web-среду.

Результатом переосмысления роли и значения библиографической информации стала концепция FRBR (Functional Requirements for Bibliographic Records – Функциональные требования к БЗ). Концепция была разработана международным библиотечным сообществом под эгидой IFLA (The International Federation of Library Associations and Institutions – Международная федерация библиотечных ассоциаций и учреждений) в 1998 г. Цели разработки заключались в изменении формы каталога для облегчения поиска и доступа к ресурсам в любых средах.

FRBR – это модель связей различных объектов, представленная как концептуальная структура, позволяющая создавать БЗ независимо от различных правил каталогизации и формата представления. Разработчики концепции опирались на важнейшие пользовательские интересы: найти, идентифицировать, выбрать и получить доступ к определенному ресурсу.

FRBR включает три группы объектов:

Группа 1: произведение (work), выражение (expression), воплощение (manifestation), физическая единица (item).

Под *произведением* понимается интеллектуальная или художественная идея (абстрактное понятие). Под *выражением* понимается абстрактная реализация интеллектуальной или художественной идеи в виде текста, музыки, изображения или любой комбинации этих средств. Под *воплощением* понимается физическая реализация одного или нескольких выражений. Воплощение возникает, когда выражение фиксируется на тот или иной носитель – бумагу, кинопленку, CD-ROM, DVD и т. д., и представляет все физические объекты, характеризующиеся одинаковым интеллектуальным и художественным содержанием и одинаковой физической формой. Примером воплощения для книги является тираж (конкретный объект). Под *физической единицей* понимается конкретный экземпляр воплощения.

Таким образом, произведение реализуется посредством выражения, выражение – посредством воплощения, проявлением отдельного воплощения является конкретный ресурс (рис. 1).

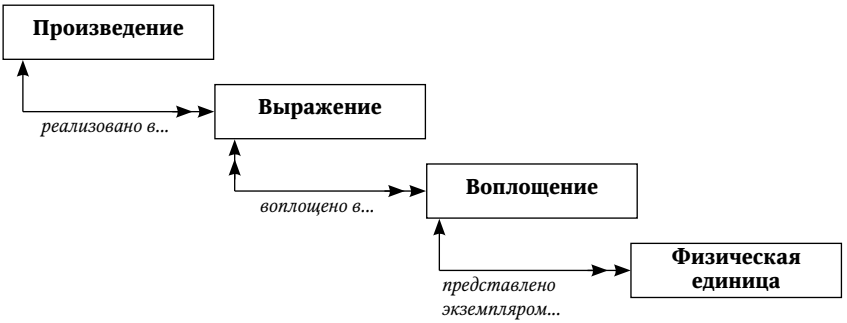

Рис. 1. Иерархические связи объектов первой группы

Группа 2: лицо (person), организация (corporate body), семья (family), несущие ответственность за интеллектуальное или художественное содержание, производство, распространение, хранение объектов группы 1 либо владеют объектами группы 1.

Группа 3: концепция/идея (concept) – абстрактное понятие или идея, предмет (object) – материальная вещь, предмет; событие (event) – действие либо период времени; место (place) – местонахождение. Эту группу составляют объекты, являющиеся предметом интеллектуальной или художественной деятельности. Кроме того, в качестве предмета произведения может выступать любой из объектов группы 1 и группы 2. Например: произведение о другом произведении или произведение об отдельной физической единице – произведение, посвященное редкой книге, существующей в единственном экземпляре.

Каждый из объектов, входящих в какую-либо группу, наделен определенными характерными чертами – атрибутами. Все объекты находятся между собой во взаимосвязях, как внутри групп, так и между объектами разных групп (рис. 2).

Связи между объектами FRBR могут использоваться для навигации пользователя в ЭК или библиографической БД[1].

---

[1] Каталогизация. Современные технологии. Тенденции и перспективы развития : курс лекций : учеб.-метод. пособие / Селиванова Ю. Г., Масхулия Т. Л., Жлобинская О. Н., Стегаева М. В. ; Рос. библ. ассоц., Рос. нац. б-ка, Нац. информ.-библ. центр ЛИБНЕТ. М. : ФАИР–ПРЕСС : Центр ЛИБНЕТ, 2007.

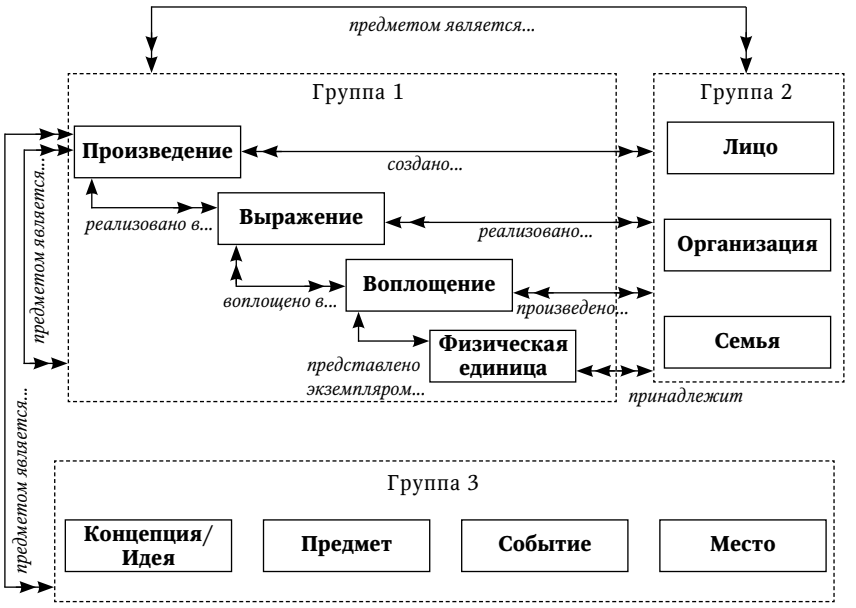

Рис. 2. Основные связи между объектами различных групп модели FRBR

Центральными связями для модели FRBR являются связи группы 1. Связи могут быть установлены между различными произведениями и их выражениями, например, между произведением А. С. Пушкина «Евгений Онегин», представленным в виде литературного текста, и музыкальным произведением П. И. Чайковского «Евгений Онегин» в виде оперы; между одним произведением и несколькими выражениями, например, текст произведения «Евгений Онегин» на русском, английском и немецком языках и т. д.

На основе формирования таких связей мы можем получить в ответ на пользовательский запрос «Евгений Онегин», например, следующий результат:

Поиск: Евгений Онегин
Записей: 49
Автор: Пушкин Александр Сергеевич (1799–1837)
Заглавие: Евгений Онегин

Найдено: (w1) Роман в стихах (текст) (45)
(w2) Опера (1)
(w3) Рецензия/Анализ (3)

Одновременно с разработкой концептуальной модели FRBR специалисты в области интернет-технологий начинают разработку новых подходов к организации эффективного поиска в возрастающем объеме информации, представленной в сети Интернет. Как известно, традиционный алгоритм поиска строится по принципу подбора ресурсов по заданным ключевым словам и не учитывает смысловой контекст информации. Для того чтобы поиск был более эффективным, поисковой машине требуется правильно интерпретировать различные данные в зависимости от запроса пользователя. Для реализации такого подхода требуется определить смысл данных и описать взаимоотношения между ними.

В результате исследований появляется концепция Semantic Web (семантическая паутина). Главным идеологом концепции Semantic Web является Т. Бернерс-Ли, один из основоположников Всемирной паутины (WWW) и директор WWW-консорциума (W3C). По определению консорциума W3C[1], Semantic Web представляет собой расширение существующей сети Интернет, в котором информация размещена в четком и определенном смысловом значении, дающем возможность обеспечить «понимание» ее компьютерами, выделение ими наиболее подходящих по тем или иным критериям данных и уже после этого – предоставление информации пользователям.

Основной акцент в концепции делается на работе с *метаданными*, однозначно характеризующими свойства и содержание ресурсов, представляемых в Web, вместо применяемого в настоящее время текстового анализа документов. В качестве метаданных могут быть использованы любые данные: даты, названия, имена, предметы и т. д., которые находятся во взаимосвязи между собой и другими данными. Для такого представления данных была разработана модель RDF (Resource Description Framework – Среда описания ресурсов), в соответствии с которой данные должны быть описаны триплетом Субъект–Отношение–Объект.

---

[1] *Хорошевский В. Ф.* Пространства знаний в сети Интернет и Semantic Web. (Часть 1). URL: http://xrumer.raai.org/library/aidt/aidt2008-1/aidt2008-1.files/2008-1-80-97.pdf.

Каждый элемент триплета должен содержать унифицированный идентификатор ресурса (URI) и иметь связь с онтологией. В примере на рис. 3 субъектом является роман в стихах «Евгений Онегин», отношением – стандарт метаданных «Дублинское ядро» (онтология), описывающий понятие «создатель», и объектом – унифицированный идентификатор создателя романа А. С. Пушкина.

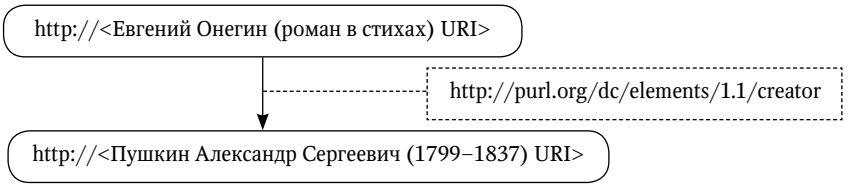

Рис. 3. Пример RDF-триплета

Таким образом, становится очевидным, что подходы к формированию и организации связи метаданных в модели FRBR и концепции Semantic Web совпадают. Это можно видеть на рис. 4.

Произведение «Евгений Онегин» в оригинале написано на русском языке, имеет выражение в виде оригинального текста и перевода на английский язык, которые изданы в виде книг разными издательствами в разные годы. Все перечисленные объекты имеют собственные URI и связаны между собой отношениями, показанными на рис. 4 стрелками. Также существует, например, опера по мотивам произведения, которая имеет собственное выражение в виде аудио- и видеозаписи. Соответственно, на уровне произведения можно установить ассоциативную связь.

Таким образом, можно говорить о том, что инструменты Semantic Web позволят библиотекам наиболее полно реализовать концепцию FRBR и формировать свои данные как часть Всемирной паутины.

В настоящее время концепция Semantic Web из стадии исследовательского проекта переходит в область практической реализации. Крупные компании, такие как Microsoft, IBM, Adobe, Sun Microsystems, Google и др., активно используют технологию Semantic Web в своих продуктах для решения задач управления данными.

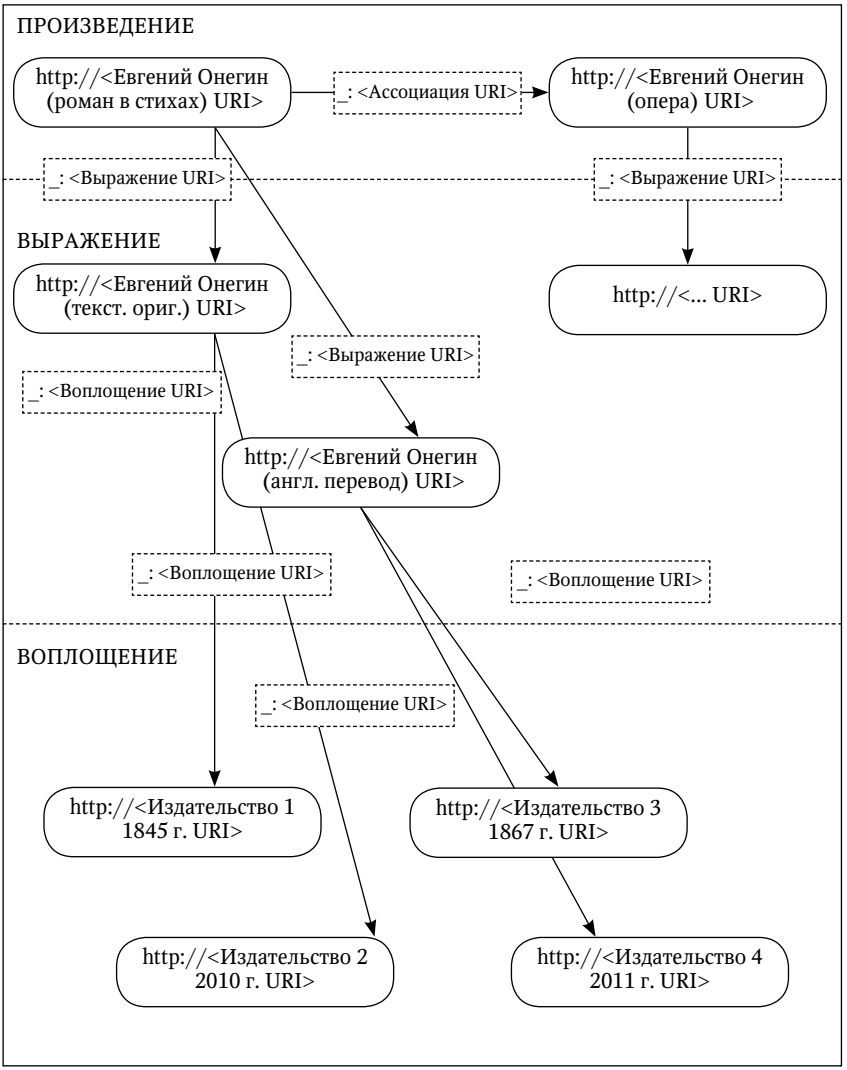

Рис. 4. Пример представления модели FRBR в структуре RDF (модель FRBR и семантический Web)

Библиотечное сообщество также находится на этапе перехода от теории к практике. На основе концептуальной модели FRBR американскими библиотечными специалистами разработан новый стандарт описания ресурсов – RDA (Resource Description and Access – Описание ресурсов и доступ к ним). RDA содержит набор правил для приведения описательных данных об опознаваемом информационном ресурсе, осязаемом или неосязаемом, т. е. книге, журнале, звукозаписи, изображении (движущемся; неподвижном; двухмерном; трехмерном; визуальном; тактильном и т. д.) и об объекте, связанном с ресурсом, т. е. лице, семье, организации, предмете, месте, событии и т. д. В 2008–2011 гг. в библиотеках США проводилось тестирование нового стандарта описания. По результатам тестирования было принято решение о внедрении нового стандарта описания в американских библиотеках, которое предполагается начать в марте 2013 г.

В качестве новой структуры представления информации о библиотечных ресурсах предполагается использовать структуру описания RDF, являющуюся базовой структурой описания данных в Semantic Web. Уже сейчас ряд библиотек мира работает над переводом своих ЭК, представленных в традиционных MARC-форматах, в структуры RDF. Наиболее масштабный проект – это представление Британской национальной библиографии в RDF-структурах. В рамках соответствующего проекта было конвертировано около 3 млн библиографических записей (см. рис. 5).

Библиографические данные описаны с использованием следующих словарей: Bibliographic Ontology, Vocabulary for Biographical Information, British Library Terms, Dublin Core, Event Ontology, Friend of a Friend, ISBD, Organization Ontology, OWL, SKOS, RDF Schema, WGS84 Geo Positioning и др. Данные опубликованы в форме открытых связанных данных (LOD Open Linked Data), что означает их свободное использование. Установлены связи со следующими ресурсами, представленными как открытые данные: VIAF, LCSH, Lexvo, GeoNames, MARC, Dewey.info, RDF Book Mashup.

Все элементы Semantic Web должны быть зарегистрированы и описаны с помощью онтологий для обеспечения технической и смысловой интероперабельности, понимаемой как способность к взаимодействию различных систем. Регистрация словарей осуществляется на специализированных сайтах, например: http://dublincore.org/; http://metadataregistry.org/ (см. рис. 6).

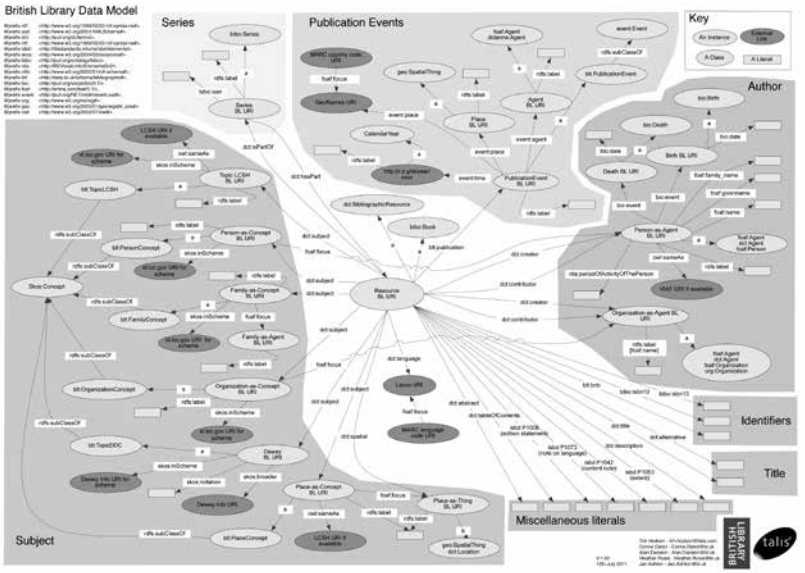

Рис. 5. Модель данных для книг
Британской национальной библиографии[1]

Отметим, что одним из важнейших элементов Semantic Web являются онтологии, включающие в том числе словари контролируемой лексики. Библиотеки обладают огромным опытом в области формирования и ведения подобных словарей в форме авторитетных файлов имен лиц, наименований организаций, географических названий, предметов. Данные словари могут широко использоваться и за пределами библиотечного сообщества. Например, Библиотека Конгресса США начиная с 2006 г. приступила к представлению своих авторитетных данных в RDF SKOS. В этой связи служба поддержки авторитетных файлов Библиотеки Конгресса США была переименована и теперь называется «LC Linked Data Service Authorities and Vocabularies» (Служба связанных данных Библиотеки Конгресса, авторитетные файлы и словари) (см. рис. 7). Есть и другие подобные примеры.

[1] URL: http://www.bl.uk/bibliographic/pdfs/bldatamodelserial.pdf.

Рис. 6. Примеры Web-сайтов для регистрации словарей

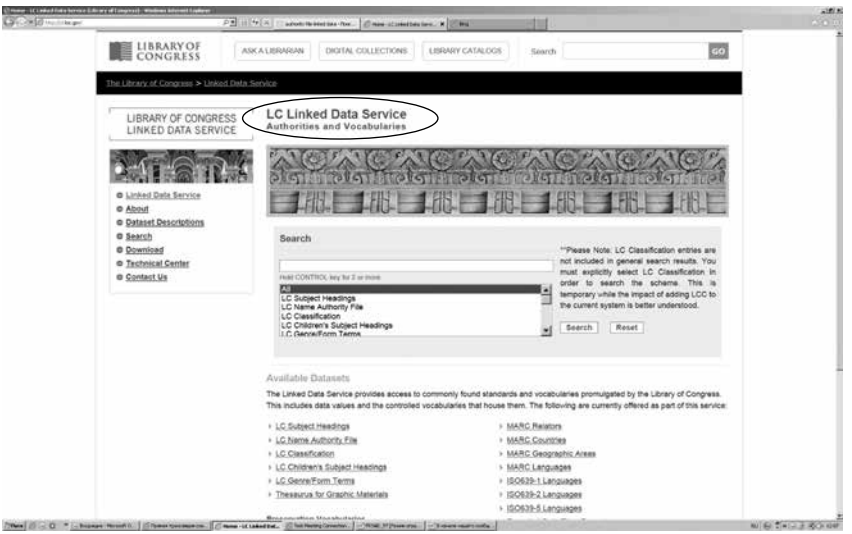

Рис. 7. Страница Библиотеки Конгресса США с авторитетными файлами

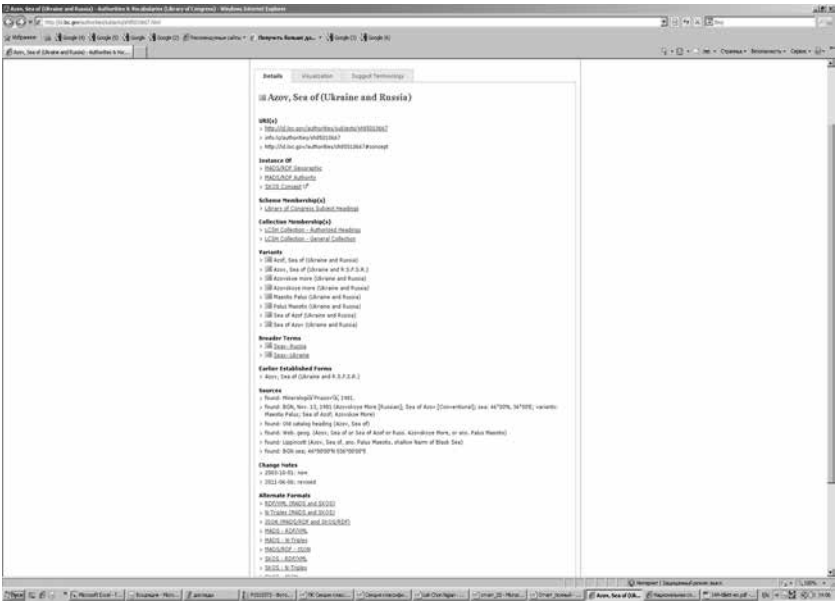

Рис. 8. Авторитетная запись предметной рубрики из LCSH в пользовательском просмотре[1]

На рис. 8 приведен пример предметной рубрики «Азовское море». Информация о ней может быть доступна в любом из форматов Semantic Web. На рис. 9 представлен фрагмент записи в формате SKOS-N-Triples. Зарубежные специалисты, осуществляющие подобные проекты, отмечают, что технических проблем с переводом библиотечных данных в структуры RDF не возникает[2].

С целью еще большего вовлечения библиотек в процесс формирования метаданных в соответствии с технологиями Semantic Web Консорциумом WWW в 2010 г. была организована Рабочая группа, включающая специалистов в области Semantic Web и библиотечных специалистов. Задача Рабочей группы заключалась в подготовке рекомендаций по представлению библиотечных данных в Web-среде. В 2011 г. Рабочая группа подготовила отчет по библиотечным

[1] URL: http://id.loc.gov/authorities/subjects/sh85010667.html.

[2] URL: http://www.bl.uk/bibliographic/pdfs/bldatamodelserial.pdf.

<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#prefLabel> "Azov, Sea of (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2008/05/skos-xl#altLabel> _:bnode7authoritiessubjectssh85010667 .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2008/05/skos-xl#altLabel> _:bnode12authoritiessubjectssh85010667 .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Azof, Sea of (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Azov, Sea of (Ukraine and R.S.F.S.R.)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Azovskoe more (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Azovskoye more (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Maeotis Palus (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Palus Maeotis (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Sea of Azof (Ukraine and Russia)"@en .
<http://id.loc.gov/authorities/subjects/sh85010667> <http://www.w3.org/2004/02/skos/core#altLabel> "Sea of Azov (Ukraine and Russia)"@en .
...

Рис. 9. Фрагмент записи предметной рубрики в SKOS-N-Triples Библиотеки Конгресса США

связанным данным *(Library Linked Data Incubator Group Final Report, W3C Incubator Group Report 25 October 2011)*[1], включивший рекомендации по использованию принципов Semantic Web и Linked Data для обеспечения процесса создания и хранения информации, представленной в виде библиографических данных, авторитетных данных, тематических списков, с целью обеспечения их восприятия и доступности в Web.

В конце 1990-х гг. российское библиотечное сообщество отставало от зарубежного сообщества по развитию и внедрению новых информационных технологий в среднем на 15–20 лет. В то время как

[1] URL: http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/.

за рубежом уже задумывались о новых концепциях представления метаданных, в России библиотеки только начинали массовое внедрение ЭК на MARC-форматах и перевод контролируемых справочников в машиночитаемую форму. К настоящему времени российские библиотеки уже обладают значительными массивами метаданных и контролируемых словарей, переведенных в MARC-форматы. Так, например, в настоящее время объем СКБР составляет около 6 млн библиографических записей на ресурсы, хранящиеся в библиотеках России, и около 2 млн авторитетных записей. Однако эта информация недоступна для широкого круга пользователей.

С целью обеспечения свободного доступа к информации российскому библиотечному сообществу необходимо воспользоваться опытом зарубежных коллег по представлению библиотечных данных в Web. Социальная значимость и масштабность задачи по переводу библиографических данных с использованием новой модели RDF в Web-среде требует поддержки на государственном уровне. Представляется, что на первом этапе необходимо создание Рабочей группы, включающей специалистов в области библиотечного дела и информационных технологий, для разработки программы действий по интеграции российских библиотек в Semantic Web.