УДК 004.94
DOI: 10.15827/0236-235X.116.036-044
Дата подачи статьи: 28.06.16
2016. Т. 29. № 4. С. 36–44

ИНФОРМАЦИОННАЯ МОДЕЛЬ СЕМАНТИЧЕСКОЙ БИБЛИОТЕКИ LibMeta

О.М. Атаева, младший научный сотрудник, oli@ultimeta.ru
(Вычислительный центр им. А.А. Дородницына Федерального исследовательского центра «Информатика и управление» РАН, ул. Вавилова, 40, г. Москва, 119333, Россия)

В данной статье библиотеки рассматриваются как информационные системы, обеспечивающие основную функциональность для работы с библиотечными данными. Развитие технологий и возможностей сети переопределяет понятие как самих библиотек, так и ее ресурсов, которые сегодня не ограничиваются только библиографическими записями и их электронным представлением, а выводят на передний план семантику этих ресурсов.

Благодаря развитию технологий пользователь библиотеки получает дополнительные возможности для работы с ресурсами цифровых библиотек с помощью описания области своих интересов в терминах предметной области на основе стандартов с привлечением словарей, тезаурусов и онтологий. Это позволяет ему организовывать и описывать и собственные коллекции, и собственные ресурсы, при необходимости детализируя как описания ресурсов, так и область своих интересов путем уточнения ее терминов.

В работе рассматриваются основные требования к таким библиотекам и описывается информационная модель разрабатываемой системы, особенностью которой является возможность интеграции данных из источников, интегрированных в облако LOD.

***Ключевые слова**: семантические библиотеки, LOD, информационная модель, интеграция данных.*

В данной статье, говоря о библиотеках, будем иметь в виду информационные системы, обеспечивающие основную функциональность для работы с библиотечными данными. Обычно при описании библиотек предполагают, что ее ресурсы представляют собой библиографические записи традиционных библиотек и электронные копии документов, описываемых этими записями. Но развитие технологий и возможностей сети переопределяет понятие как самих библиотек, так и ее ресурсов, которые не ограничиваются теперь только библиографическими записями и их электронным представлением, а также выводит на передний план семантику этих ресурсов. Для этого разработаны различные виды классификации ресурсов библиотеки – отраслевые рубрикаторы, позволяющие более детально определять тематическую направленность ресурсов. Зачастую этих средств для описания семантики недостаточно, со временем появляются новые требования к описанию ресурсов библиотек, что приводит к усложнению самих описаний и требует значительных трудозатрат на внедрение новых способов описаний, соответствующих текущим потребностям. Увеличивающийся поток поступающих объектов практически невозможно обработать вручную, нужны новые методы обработки и анализа поступающих данных.

В современных библиотеках сами ресурсы становятся более разнообразными и могут включать самые разнотипные объекты. Например, электронная библиотека «Научное Наследие России» [1], заявленная как проект по созданию библиотеки полнотекстовых научных трудов известных российских и зарубежных ученых и исследователей, включает в себя и описания музейных экспонатов, расширяя тем самым традиционные типы хранимых ресурсов классической библиотеки.

Semantic Web, Linked Open Data и библиотеки

В последнее десятилетие возросла популярность парадигмы Semantic Web [2], одним из практических воплощений которой стало сообщество, поддерживающее публикацию данных в сети согласно принципам LOD (Linked Open Data) [3]. Основным преимуществом этого подхода является возможность провязывания ресурсов из различных источников данных, при описании которых используются онтологии, содержащие метаописания самих метаданных ресурсов.

Основная идея LOD заключается в решении задач интеграции данных сети, для чего предлагается представлять информацию в формализованном виде, что делает ее доступной для машинной обработки. Единицей описываемых данных в Semantic Web является ресурс. Каждый ресурс обозначает какой-либо реальный объект, понятие или явление и имеет идентификатор URI (Unified Resource Identificator), который используется для описания знаний о сущности. Эти знания представляются в соответствии с моделью данных RDF (Resource Definition Framework) [4] в виде троек «субъект–предикат–объект».

Организация специального пространства связанных данных Linked Data основывается на практических решениях для публикации и связывания структурированных данных. Термин LOD описывает ту часть данных Linked Data, которая находится в открытом доступе и соответствует основным принципам LOD. Идея LOD очень привлекательна для различных организаций, многие из которых включили свои источники данных в это облако. Оказались провязанными самые различные типы ресурсов, которые представляют интерес для

пользователей библиотек с точки зрения обогащения данных как структурно, так и семантически.

Главные проблемы уже существующих наборов данных в LOD на текущий момент – это разнообразие терминов и разобщенность данных. В разных наборах данных могут использоваться различные онтологии для описания модели данных. Классический случай – библиографические онтологии, описывающие модель данных для ведения библиографических записей печатных изданий. Часто встречаются библиотеки, контент которых – это набор тематических ресурсов, для их поддержки составляется соответствующая онтология. Например, в качестве ресурсов могут выступать некоторые мультимедийные объекты, для описания которых классические библиографические онтологии, такие как BIBO или SPAR, непригодны. Эти проблемы являются следствием нестандартизованности процесса публикации набора данных в пространство LOD, а также иллюстрируют важность тщательного выполнения интеграции и реализации возможностей семантического поиска.

В связи с развитием технологий пользователь библиотеки может получать больше возможностей для работы с ресурсами цифровых библиотек с помощью описания области своих интересов в терминах предметной области на основе стандартов с привлечением словарей, тезаурусов и онтологий. Это позволит ему организовывать и описывать собственные коллекции и собственные ресурсы, при необходимости детализируя как описания ресурсов, так и область своих интересов посредством уточнения ее терминов.

Определения библиотек

Формально ***электронная библиотека*** представляет собой тройку объектов $<F, C, A>$, где F – множество функций хранения и поиска, обеспечиваемых информационной системой для обработки объектов множества C. Объекты из C обладают фиксированным набором атрибутов $(a_1, \ldots, a_k)$, $a_i \in A$. Этот набор будем называть *описанием множества C* или *метаданными множества C*.

Множество C будем называть контентом библиотеки, а любой объект $c \in C$ – информационным объектом. Тогда описание отдельного информационного объекта обозначим как $c(a_1, \ldots, a_k)$. При этом значениями a_i могут быть только символьные наборы из некоторого алфавита L. Набор атрибутов и символьные значения этих атрибутов для объекта $c \in C$ – метаданные этого объекта. Значения атрибута a_i обозначим $c(a_i) \in L^*$, где L^* – множество всех строк (включая пустую строку), составленных из символов, входящих в L. Множество F состоит из функций вида f: $(a_1, \ldots, a_j) \rightarrow In$, $In \subset C$, j принимает значения от 1 до k.

Фактически контент электронных библиотек представляет собой множество библиографических записей объектов реальной классической библиотеки. В электронных библиотеках речь идет не о цифровом представлении копий реальных объектов, а лишь об их описаниях. В таких описаниях, например, встретив упоминание персоны в одном месте, невозможно точно установить соответствие с ее упоминанием в другом месте.

Цифровые библиотеки решают те же задачи поиска и хранения контента, что и электронные библиотеки, но существенно расширяют свою функциональность и определение своего контента. Во-первых, контент библиотеки становится мультимедийным. Это значит, что значениями атрибутов ее информационных объектов теперь могут выступать различные мультимедийные объекты, доступные для просмотра средствами самой цифровой библиотеки. В качестве мультимедийных объектов могут выступать совокупность аудио-, видео-, фото- и текстовых материалов. Во-вторых, расширяется функциональность за счет решения некоторых задач интеграции как метаданных, так и медийных объектов из внешних источников, доступных по сети.

При этом формальное определение представляет собой уже набор объектов $<F, C, A, M>$, где F, C, A определяются так же, как и в электронных библиотеках; M – множество доступных мультимедийных объектов, $c(a_i) \in (M \cup L^*)$.

Множество F дополняется функциями вида g: $(X, a_1, \ldots, a_j) \rightarrow Out$, где $Out \subset C(X)$ и $Out \subset C$, X – внешний источник, а множество объектов $C(X)$ может быть описано набором атрибутов $(a_1, \ldots, a_k)$. Функции g предназначены для решения вопросов интеграции данных из внешних библиотек.

Семантические цифровые библиотеки являются следующим этапом в эволюции библиотек и обязаны своей популярностью семантическим технологиям, которые в значительной степени повлияли на переосмысление понятия контента библиотеки и послужили толчком для расширения и улучшения функциональности библиотек. В таких библиотеках данные лучше структурированы, выделены связи между ними, что улучшает поиск и дает возможность интегрировать данные различных типов. Обеспечивается интероперабельность с другими системами, необязательно являющимися библиотеками, так как основной задачей семантических технологий является предоставление метаданных в машиночитаемом формате.

Формально семантическая цифровая библиотека – это $<F, C, A, \Phi, T>$, где F и A определяются так же, как и в цифровых библиотеках. Контент библиотеки $C = C_1 \cup C_2 \cup \ldots \cup C_s$ представляет собой множество типов информационных ресурсов системы, для каждого их которых определен свой набор атрибутов $(a_{i1}, \ldots, a_{ik})$. Такое определение не означает исключение мультимедийных объектов, а подчеркивает обыденность мультимедийных объектов в семантических библиотеках, то есть $M \in C$,

и должно пониматься как добавление нового типа контента *«мультимедийный объект»* в библиотеку со своим набором атрибутов и отношений, каждый объект которого является абстрактным представлением реального объекта из множества M. Значения атрибутов $c(a_{ij}) \in (M \cup L^* \cup C)$; L^*, как и прежде, содержит область значений строковых атрибутов из A. Φ задает множество условий, накладываемых на представление контента, которое может, например, содержать ограничения, накладываемые на форматы значений $c(a_{ij})$. T – множество терминов предметной области, предназначенных для терминологической поддержки описания экземпляров информационных ресурсов множества C. Объектом T может быть подмножество элементов L^* или множество подмножеств из L^*. Элементы множества T могут быть связаны различными отношениями между собой, образуя простые таксономии (линейный словарь, классификатор с иерархическими связями) или сложные таксономии (таксономии с горизонтальными связями), а также могут быть связаны отношениями с объектами из множества C.

Важную роль в определении семантических библиотек при описании ее контента играют онтологии. Онтология модели контента фактически задается $<C, A, \Phi, T>$, где множество C выступает как множество понятий онтологии, множество атрибутов A также содержит подмножество отношений между понятиями, а Φ задает множество функций интерпретации, заданных на понятиях и отношениях. Таким образом, множества C, A, Φ задают описание структуры контента библиотеки, тогда как объекты множества T терминологически ограничивают предметную область контента библиотеки.

Основные свойства семантических библиотек

Основным свойством семантических библиотек является возможность структурирования их разнообразного контента и связывания данных из разных источников, что, в свою очередь, отражается на качестве данных контента.

Основные свойства семантической библиотеки, которые, на взгляд автора, являются определяющими для рассматриваемой системы:

– семантическая библиотека представляет собой интеграционный узел для разных источников данных, которые обогащают и пополняют ее набор данных;

– контент библиотеки описывается на семантическом уровне, что позволяет достичь лучшего взаимодействия между источниками данных;

– контент библиотеки может иметь разную степень гранулярности структуры в зависимости от рассматриваемых начальных условий при построении библиотеки;

– семантическое описание контента и его уровень гранулярности не зависят от технических характеристик реализации информационной системы библиотеки;

– понятийное описание контента библиотеки поддерживается его тезаурусом, который ограничивает предметную область ресурсов библиотеки терминологически.

Информационные системы в контексте семантических библиотек

Выделив модель контента семантической библиотеки и ее основные характеристики, отделим определяющее понятие контента семантической библиотеки от понятия реализующей библиотеку информационной системы [5]. Такой подход позволяет наращивать функциональность системы, добавлять новые подсистемы или изменять уже имеющиеся при неизменных остальных частях.

Информационная система IS задается набором подсистем F для решения задач обработки ее контента I, $IS = (F, I)$. Тогда IS представляется как организация совокупности своих подсистем $F = \cup F_i$ и своего контента I. Каждая из этих подсистем описывается своей предметной онтологией, и тогда можно представить ***онтологию информационной системы*** $OnIS$ объединением онтологий ее подсистем и онтологии ее контента $OnIS = OnF \cup OnI$, где $OnF = \cup OnF_i$ – объединение онтологий подсистем, $OnI = <C, A, \Phi, T>$ – онтология контента. При описании онтологий информационных систем и ее модулей обычно опираются на абстрактные онтологии высокого уровня, определяя ее ключевые сущности. Более подробно этот подход освещается в [6].

Основные виды задач реализуются в информационной системе подсистемами:

– описания контента информационной системы;

– реализации задач интеграции данных из внешних источников;

– реализации задач интеграции данных из внутренних источников;

– поддержки коллекций;

– поиска и навигации по объектам системы;

– поддержки пользователей;

– управления тезаурусом;

– качества данных в системе.

Такое разбиение на подсистемы не является единственно возможным. Границы подсистем не могут быть строго определены. В системе существует область общих определяющих понятий, которые рассматриваются как принадлежащие нескольким подсистемам, в зависимости от того, какие процессы выполняются в конкретной подсистеме. Каждая из этих систем так или иначе взаимодействует с понятиями, определяющими контент этой библиотеки. Например, в перечисленных

подсистемах можно рассматривать как единую подсистему реализации задач интеграции данных из внешних и внутренних источников. С другой стороны, из подсистемы качества можно выделить отдельно систему выявления дубликатов. Это деление диктуется конкретной реализацией.

Краткий обзор некоторых семантических библиотек

Семантическая библиотека JeromeDL [7] является одной из попыток объединить возможности, предлагаемые концепцией и технологиями Semantic Web, с библиотеками, ориентируясь на тесное взаимодействие с пользователями. Фактически она представляет собой интегрированное приложение для ведения цифровой библиотеки, блогов и сервиса для закладок. В рамках цифровой библиотеки поддерживаются авторитетные файлы (для авторов, редакторов, издательств), таксономии, используемые для классификации по темам, тезаурус WordNet [8] для ключевых слов. Каждый ресурс описывается тремя типами метаданных: структурными, библиографическими и социальными. Каждый тип метаданных поддерживается соответствующими сервисами. Пользователю предоставляется комбинированное представление на основе этих метаданных. Основные модели для описания ресурсов, пользователей и их взаимодействия – библиографическая онтология MarcOnt [9], онтология FOAF [10], модель знаний SKOS [11] для описания таксономий.

Основными недостатками, на взгляд автора, являются ориентированность только на библиографические данные и слабая поддержка интеграции данных с другими источниками в рамках системы, в частности, с ресурсами из LOD. При необходимости добавления нового типа ресурсов приходится вносить изменения в систему на программном уровне. Одним из преимуществ этой системы является поддержка, помимо сервиса традиционного атрибутного поиска, сервисов семантического поиска данных на естественном языке, доступ к данным на языке запросов SPARQL [12] для возможности машинной обработки. Следует отметить, что система распространяется бесплатно.

Остальные решения в этой области, такие как Greenstone [13] и Briks [14], так или иначе уступают указанному проекту или предназначены для использования в рамках специализированных предметных областей.

Одной из глобальных реализованных цифровых библиотек является проект Europeana [15], интегрирующий данные из институтов культурного наследия Европы. Многоуровневая организация провайдеров контента предназначена для автоматической оценки контента на соответствие модели данных EDM (Europeana Data Model) [16], которая была разработана в рамках проекта. В рамках этой модели данных определены наборы классов и свойств для описания объектов культурного наследия. Одно из преимуществ EDM – возможность соблюдения принципов связанных данных при описании ресурсов. Масштаб этой библиотеки одновременно является и одним из препятствий для возможности индивидуальной тематической работы пользователя и позволяет причислить ее к глобальным семантическим библиотекам, среди которых также можно указать DBpedia [17], являющуюся ядром облака LOD.

Информационная модель цифровой библиотеки LibMeta

Наиболее полной эталонной моделью электронной библиотеки является разработка группы DELOS (Digital Library Reference Model, DLRM) [18]. Были определены базовые для электронной библиотеки понятия (конкретная ЭБ, система ЭБ, система управления ЭБ), выделены категории пользователей для этих понятий (разработчик, пользователь, администратор) и шесть основных высокоуровневых понятий/областей: контент, пользователь, функциональные возможности, качество, политики, архитектура.

Опираясь на концептуальную модель DELOS и ее определения, а также на идеи Semantic Web и Linked Open Data, была разработана персональная открытая семантическая библиотека LibMeta с системой поддержки работы пользователей с цифровыми ресурсами библиотек и их коллекциями для некоторой предметной области.

Основные требования к системе LibMeta. LibMeta характеризуется настраиваемым хранилищем метаданных для своих ресурсов и типами описываемых информационных ресурсов. Основные требования к описанию ресурсов – универсальность, структурированность и адаптируемость. Универсальность – независимость описания ее типов ресурсов и объектов от предметной области и области интересов пользователей. Структурированность описания обеспечивает поддержку связей между различными типами ресурсов как внутри системы, так и вне ее, исходя из определений LOD. Адаптируемость описания ресурсов обеспечивает возможность добавления новых свойств и связей в процессе развития системы и обеспечивает настройку пользовательских интерфейсов под эти изменения. Далее приведены основные понятия подсистем LibMeta, которые обеспечивают соответствие этим требованиям, выведенным на базе формальной модели семантической библиотеки [19].

Подсистема описания контента информационной системы. За универсальность определения контента системы отвечает набор понятий, составляющих информационную модель контента библиотеки LibMeta: информационный ресурс и

информационный объект, которые описывают экземпляры ресурсов. Информационный ресурс является основной единицей описания контента библиотеки, а информационный объект представляет экземпляры информационных ресурсов. Каждый из них имеет собственный уникальный идентификатор в соответствии с требованиями LOD. Фактически семантическое значение информационного ресурса является эквивалентным понятию класса онтологии с некоторыми ограничениями в его описании. Структура описания информационных объектов определяется понятиями *атрибут* и *набор атрибутов*, которые задаются при описании соответствующего ресурса. Атрибут является элементом описания свойства ресурса, а набор атрибутов – коллекцией атрибутов разных видов. Типы атрибутов следующие: атрибут, файловый, объектный, числовой, текстовый, строковый. При подключении подсистемы управления таксономиями появляется новый вид атрибута – таксономический. Помимо определения круга значений атрибута, важными характеристиками являются тип и количество его значений [19].

Эти понятия обеспечивают структурированное описание контента и обеспечивают поддержку его адаптируемости. Такой подход также обеспечивает описание конкретных ресурсов и их объектов в виде RDF-троек и предоставления SPARQL точки доступа для публикации данных в LOD.

Конкретная реализация модели контента библиотеки может быть основана на некоторой импортируемой онтологии, классы которой превращаются в ресурсы, свойства описываются в терминах атрибутов LibMeta, наборы атрибутов определяют фактически домены свойств онтологий. При построении модели ресурсов библиотеки на основе этой онтологии сохраняются все URI свойств, отношений и классов выбранной онтологии. При необходимости при импортировании выбранной онтологии в систему можно изменить набор понятий, расширив или сократив его средствами системы.

Конечно, такой способ отображения онтологии на понятия системы LibMeta не сохраняет весь возможный перечень ограничений, накладываемых на свойства и классы онтологии изначально, но ее структурная часть сохраняется, что является достаточным для решения задач, определенных в рамках системы.

На рисунке 1 приведены основные понятия, используемые для конструирования описания предметной области. Некоторые из понятий будут пояснены далее.

Подсистема управления тезаурусом. Для описания *тезауруса* введены дополнительные понятия: *таксон* и *таксономия*. Таксон представляет собой элемент таксономии с определенным набором свойств, необходимым для его базового представления, а таксономия определяет набор доступных связей между составляющими таксономию таксонами и ресурсами системы. Для описания дополнительных связей между таксонами вводятся отношения между ними, которые позволяют определять и описывать новые связи в рамках информационной системы. По умолчанию в системе доступны только два типа связей между таксонами: иерархическая и нетипизированная горизонтальная. На рисунке 2 представлены поддерживаемые подпонятия таксономии согласно определению семантической библиотеки: словарь, классификатор и тезаурус. На рисунке 3 отображаются используемые по умолчанию связи в таксономиях между определяющими их таксонами. Для тезауруса можно доопределить атрибуты, помеченные на ри-

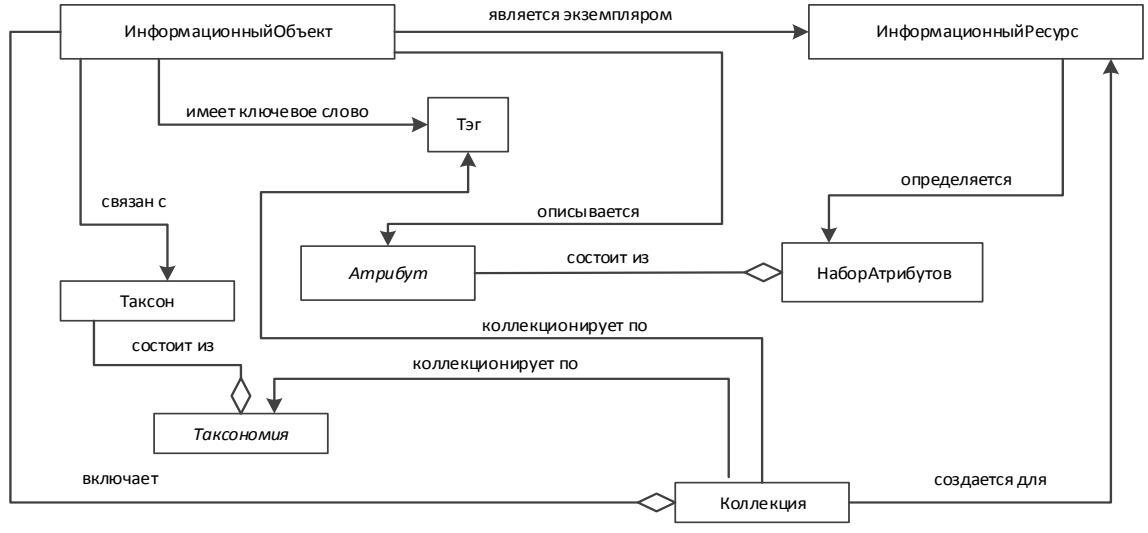

Рис. 1. Основные понятия и связи между ними

Fig. 1. Basic concepts and their relationships

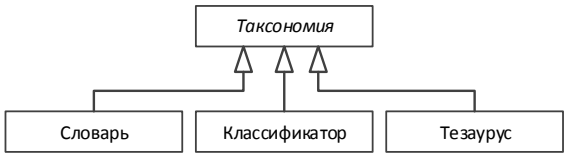

Рис. 2. Таксономии

Fig. 2. Taxonomies

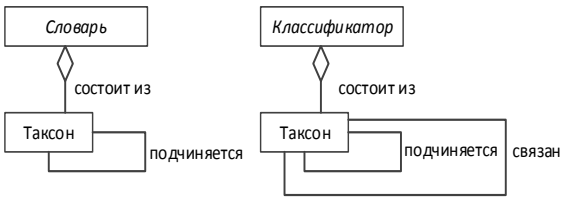

Рис. 3. Связь таксономий и таксонов

Fig. 3. Relations between taxonomies and taxons

сунке 4 как *T_ТекстовыйАтрибут, T_ОбъектныйАтрибут, T_ТаксономическийАтрибут*, которые позволяют расширить определение таксона, включать в его определение также информационные объекты и специфировать горизонтальные связи между таксонами.

Для определения связи таксономий с информационными объектами вводится понятие *таксономического атрибута*. Это отношение обеспечивает возможность подключения любой из таксономий к любому типу ресурсов в процессе жизнедеятельности системы. Такой подход позволяет, с одной стороны, избежать избыточности на начальном этапе проектирования системы, с другой – обеспечить представление практически любых связей. Таксономические атрибуты задаются при описании области значений атрибутов информационных ресурсов.

Подсистема поддержки коллекций. Для возможности ведения разнообразных коллекций объектов используется понятие *коллекция информационных объектов*, которая определяется на основе некоторой таксономии с указанием коллекционируемых типов ресурсов. Коллекция может объединять информационные объекты различных информационных ресурсов. На основе одной и той же таксономии можно определять несколько коллекций. Такой подход окажется чрезвычайно полезным для создания раздельных пользовательских коллекций.

Подсистема реализации задач интеграции данных из внешних источников. Для решения задач интеграции данных из источников LOD вводится понятие *источник данных*, которому ставятся в соответствие информационные ресурсы системы, и устанавливается соотношение набора атрибутов ресурса со свойствами ресурса из источника данных. Это позволяет генерировать SPARQL-запросы к источникам данных для извлечения конкретной информации. При этом пользователь оперирует привычными формами поиска, избегая необходимости написания самих запросов.

Для случая, когда конкретная реализация модели контента библиотеки основана на некоторой импортируемой онтологии и онтология используется в источнике данных, предусмотрен механизм взаимно однозначного отображения свойств и классов онтологий подключаемого набора данных в термины LibMeta полуавтоматическим способом. Таким образом, формируется интеграционный узел, который позволяет устанавливать взаимосвязи с источниками данных, расположенными в LOD. На рисунке 5 приведена схема связей понятия *источника данных* с основными понятиями, определяющими контент библиотеки. Рисунок 6 иллюстрирует взаимодействие пользователя с подсистемой для получения результатов своего запроса.

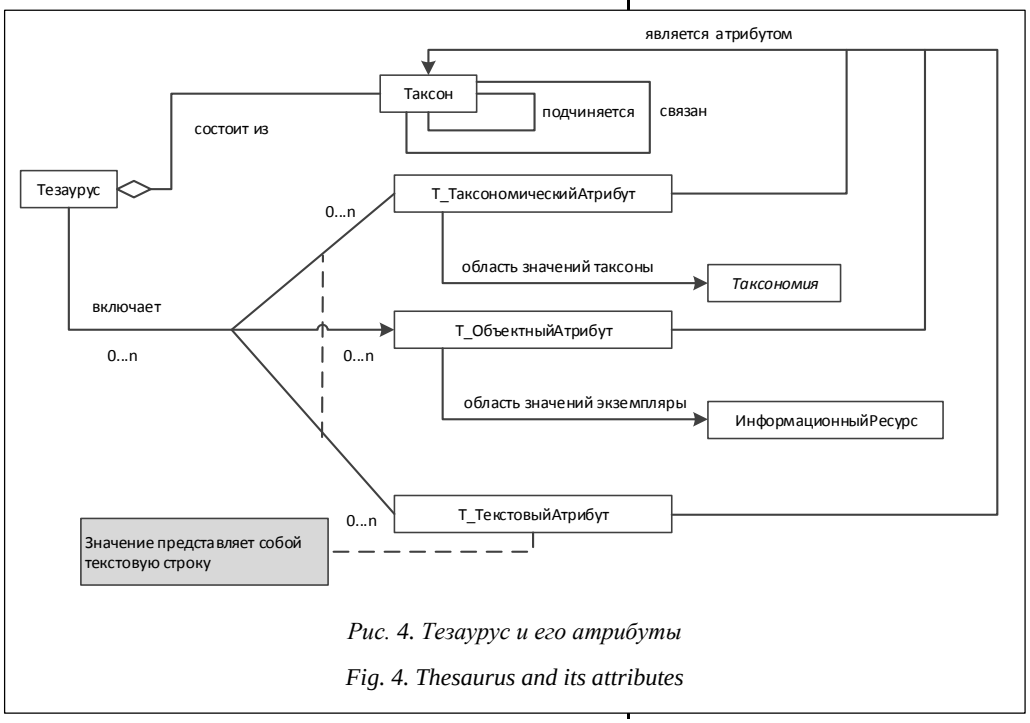

Рис. 4. Тезаурус и его атрибуты

Fig. 4. Thesaurus and its attributes

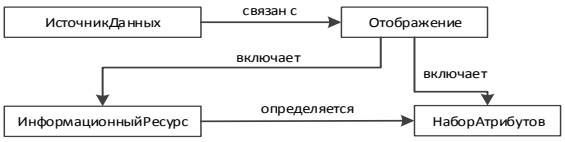

Рис. 5. Источник данных и основные понятия контента

Fig. 5. The data source and the basic concepts of content

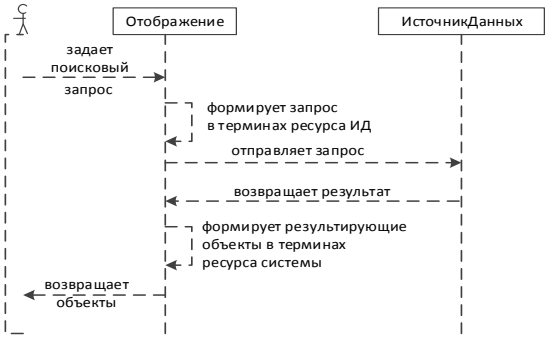

Рис. 6. Взаимодействие пользователя с источником данных

Fig. 6. User interaction with the data source

Подсистема реализации задач интеграции данных из внутренних источников. На самом деле эта подсистема оперирует теми же понятиями, что и предыдущая: *источник данных, ресурс, атрибуты ресурса*. Главное отличие состоит том, что интегрируемый источник не включен в LOD. Доступ к таким источникам обычно осуществляется по некоторому протоколу, который может быть широко используемым, таким как OAI-PMH, и довольно специфичным. Для таких случаев предусмотрена отдельная подсистема, которая поддерживает извлечение данных практически по любым протоколам и возвращает данные, готовые для включения в систему, в терминах информационных ресурсов системы.

Подсистема поиска и навигации по объектам системы. При описании информационных ресурсов и определении набора их атрибутов важную роль играют виды атрибутов, формирующие структурное описание ресурса. Атрибуты делятся на несколько пересекающихся видов: поисковые, описательные, административные, идентифицирующие. В подсистеме поиска важную роль играют именно поисковые атрибуты, используемые при выполнении атрибутного поиска по типам ресурсов. Результатом такого поиска являются объекты, краткое описание которых представлено пользователю посредством описательных атрибутов.

Подсистема поиска также поддерживает поиск по ключевым словам (или тегам) по всем ресурсам системы. Поиск выполняется по ключевым словам, которыми снабжены объекты системы.

Итак, основные понятия в этой подсистеме – *ресурс, объект, атрибут, вид атрибута, тег.* Тег объекта фактически представляет собой ключевое слово и обеспечивает функциональность определения набора семантических тегов информационных объектов из ключевых слов.

Подсистема качества данных в системе. Подсистема качества данных непосредственно использует понятия информационного ресурса, информационного объекта и атрибута, дополняя собственными понятиями *ошибка*, *тип ошибки*, *правило*, *условие*. С помощью этих понятий определяются процесс (поток) работ для устранения проблем в данных, условие выявления ошибки определенного типа для значений атрибута объектов некоторого типа ресурсов и задается правило устранения ошибок [20].

Подсистема поддержки пользователей LibMeta. Важной составляющей любой информационной системы являются ее пользователи. Рассмотрим в общих чертах основные понятия подсистемы поддержки пользователей: *пользователь, роль, разрешение, информационный ресурс, информационный объект, область интересов*. Для каждого пользователя уровень доступа определяется его *ролью*, определяющей набор *прав доступа* для работы с информационными ресурсами и объектами. Для каждого пользователя системы определяется область его интересов, в описании которой может быть задействован тезаурус предметной области контента библиотеки, а также список пользователей со сходным кругом интересов (рис. 7). При этом каждый пользователь может создавать свои коллекции ресурсов в рамках своих интересов, пользуясь доступными средствами соответствующих подсистем.

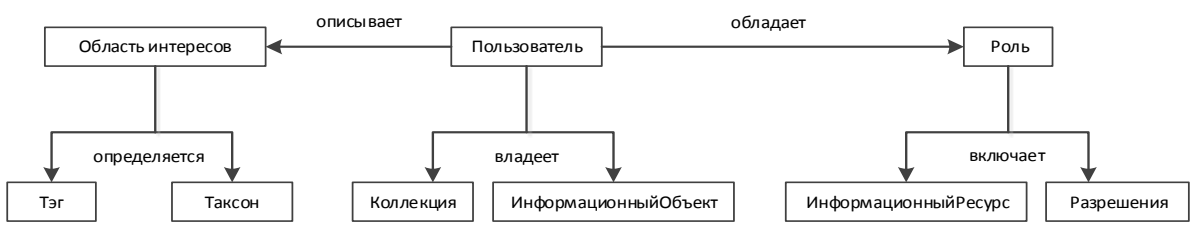

Рис. 7. Описание области интересов пользователей

Fig. 7. Description of users' interests area

Основная функциональность LibMeta. Функциональность LibMeta, доступная для всех публичных пользователей:

– просмотр ресурсов и их структуры;

– атрибутный и семантический поиск и навигация по доступным ресурсам системы;

– атрибутный и семантический поиск по источникам данных;

– просмотр общедоступных коллекций информационных объектов.

Для авторизированного пользователя LibMeta обеспечивает дополнительно следующую функциональность:

– определение своей таксономии или расширение своей ветви определенного в системе основного терминологического тезауруса; фактически обеспечивается поддержка создания так называемых *аннотационных онтологий* [21] или *онтологий пользователей* (фолксономии) [22–24], которые представляют собой коллективный словарь пользователей, составленный в результате процесса проставления тегов ими для ресурсов;

– определение собственной коллекции ресурсов, основанной на использовании таксономии;

– организация совместных тематических коллекций для групп пользователей;

– атрибутный и семантический поиск по источникам данных с возможностью сохранения результатов поиска;

– доступ администратора системы ко всей вышеопределенной и дополнительной, доступной только ему функциональности:

✓ расширение описания типов ресурсов или создание новых (по запросу пользователей);

✓ включение объектов ресурсов в общедоступный список (по запросу пользователей);

✓ обеспечение возможности редактирования определенных типов ресурсов или таксономий (для групп пользователей);

✓ редактирование групп и ролей пользователей и набора доступных им операций;

✓ редактирование и настройка основного терминологического тезауруса и его связей.

Таким образом, в статье рассмотрена информационная модель семантической библиотеки, определены основные понятия, лежащие в основе проектирования персональной открытой семантической библиотеки LibMeta. Рассмотрены общая архитектура семантической библиотеки и подход к интеграции источников данных.

В данный момент сконструирована библиотека на основе данных из «Научного Наследия России» и выполнено связывание с данными из источников LOD. Второй пример сконструированной библиотеки касается предметной области обыкновенных дифференциальных уравнений и публикаций по математике и также связан с данными из LOD.

Дальнейшее направление работ связано с созданием модуля автоматического тегирования документов или выделения их ключевых слов как для отдельных описаний объектов, так и для их коллекций. В качестве описаний, к примеру для публикаций, может использоваться ее полный текст. В случае объемных текстов предполагается возможность предварительного создания автореферата текстов.

Это позволит выполнять тематическую кластеризацию документов, построение иерархии ключевых слов по темам для персональных коллекций, использовать семантические теги при поиске объектов на естественном языке как в самой библиотеке, так и в подключенных источниках данных.

Литература

1. Научное наследие России. URL: http://e-heritage.ru (дата обращения: 25.06.2016).

2. Semantic Web. URL: http://www.w3.org/standards/semanticweb/ (дата обращения: 25.06.2016).

3. Bizer C., Heath T., and Berners–Lee T. Linked data – the story so far. Int. J. Semantic Web Inf. Syst., 2009, vol. 3, no. 5, pp. 1–22.

4. Resource Description Framework (RDF). URL: https://www.w3.org/RDF/ (дата обращения: 25.06.2016).

5. Серебряков В.А. Что такое семантическая цифровая библиотека // Электронные библиотеки: перспективные методы и технологии, электронные коллекции (RCDL): тр. 16-й Всерос. науч. конф. Дубна: Изд-во ОИЯИ, 2014. С. 21–25.

6. Palano R., Pandurino A., Guido A.L. Conceptual design of web application families: the BWW approach. Proc. 6th Workshop on Domain Specific Modeling, Portland, USA, 2006, pp. 23–32.

7. Kruk S.R. et al. JeromeDL-a semantic digital library. 2007, 268 p.

8. Miller G.A. WordNet: a lexical database for English. Communications of the ACM, 1995, vol. 38, no. 11, pp. 39–41.

9. Kruk S.R., Synak M., Zimmermann K. MarcOnt--Integration ontology for bibliographic description formats. Proc. Intern. Conf. on Dublin Core and Metadata Applications, 2005, pp. 231–234.

10. URL: http://xmlns.com/foaf/spec/ (дата обращения: 25.06.2016).

11. Isaac A., Summers E. SKOS simple knowledge organization system primer. Working Group Note, W3C. 2009. URL: https://www.w3.org/TR/skos-primer/ (дата обращения: 25.06.2016).

12. URL: https://www.w3.org/TR/sparql11-overview/ (дата обращения: 25.06.2016).

13. Witten I. H., Bainbridge D., Boddie S. J. Greenstone: Open-source digital library software with end-user collection building. Online information review. 2001, vol. 25, no. 5, pp. 288–298.

14. Aloia N., Concordia C., Meghini C. Implementing BRICKS, a Digital Library Management System. SEBD, 2007, pp. 4–15.

15. Europeana Collections. URL: http://www.europeana.eu (дата обращения: 25.06.2016).

16. Doerr M. et al. The europeana data model (edm). World Library and Information Congress: Proc. 76th IFLA General Conf. and Assembly, 2010, pp. 10–15.

17. URL: http://dbpedia.org/ (дата обращения: 25.06.2016).

18. Candela L., Castelli D., Dobreva M., Ferro N., Ioannidis Y., Katifori H., Koutrika G., Meghini C., Pagano P., Ross S., Agosti M., Schuldt H., Soergel D. The DELOS Digital Library Reference Model Foundations for Digital Libraries. IST–2002 2.3.1.12. Technology–enhanced Learning and Access to Cultural Heritage. Version 0.98, December 2007. URL: http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf (дата обращения: 25.06.2016).

19. Атаева О.М., Серебряков В.А. Основные понятия формальной модели семантических библиотек и формализация процессов интеграции в ней // Программные продукты и системы. 2015. № 4 (112). С. 180–187.

20. Атаева О.М., Бездушный А.Н. Моделирование потоков работ в задаче приведения данных // Информационное обеспечение науки. Новые технологии: сб. науч. тр. М.: Научный Мир, 2009. С. 323–329.

21. Castro L., Giraldo O.X., Castro A.G. Using the Annotation Ontology in semantic digital libraries. Proc. ISWC 2010. Posters & Demonstrations Track. Collected abstracts. Shanghai, China, Nov. 9, 2010, vol. 658, pp. 153–156. URL: http://ceur-ws.org/Vol-658/ (дата обращения: 25.06.2016).

22. Kruk S.R., McDaniel B. Semantic digital libraries. Springer, Berlin, 2009, 245 p.

23. Lee Y., Yang S.Q. Folksonomies as subject access-a survey of tagging in library online catalogs and discovery layers. IFLA Publ. Series. 2012. URL: https://www.nlib.ee/html/yritus/ifla_jarel/papers/4-1_Yan.docx (дата обращения: 25.06.2016).

24. Spiteri L.F. The structure and form of folksonomy tags: The road to the public library catalog. Information technology and libraries, 2013, vol. 26, no. 3, pp. 13–25.

Software & Systems
DOI: 10.15827/0236-235X.116.036-044

Received 28.06.16
2016, vol. 29, no. 4, pp. 36–44

AN INFORMATION MODEL OF LibMeta SEMANTIC LIBRARY

O.M. Ataeva[1], *Junior Researcher, oli@ultimeta.ru*
[1] *Institution of Russian Academy of Sciences Dorodnicyn Computing Centre of RAS, Vavilov St. 40, Moscow, 119333, Russian Federation*

Abstract. The article considers libraries as information systems that provide the core functionality for working with data. The technology development determines the concept of both the library and its resources, which are not limited only by bibliographic records and their electronic submission now, but also bring the semantics of these resources to the front.

Based on new opportunities offered by advances in technologies, a library user receives additional opportunities to work with digital library resources using descriptions of their range of interests in subject area terms based on standards with dictionaries, thesauri and ontologies. This allows organizing and describing his own collections, as well as his own resources, detailing a resource description and their area of interest by clarifying its terms.

The paper considers basic requirements for such libraries and describes the developed system information model. A feature of the system is the ability to integrate data from sources integrated in the LOD cloud.

Keywords: semantic library, LOD, data integration, information model.

References

1. *Nauchnoe nasledie Rossii* [Scientific Heritage of Russia]. Available at: http://e-heritage.ru (accessed June 25, 2016).
2. *Semantic Web*. Available at: http://www.w3.org/standards/semanticweb/ (accessed June 25, 2016).
3. Bizer C., Heath T., Berners-Lee T. Linked data – the story so far. *Int. J. Semantic Web Inf. Syst*. 2009, vol. 3, no. 5, pp. 1–22.
4. *Resource Description Framework (RDF)*. Available at: https://www.w3.org/RDF/ (accessed June 25, 2016).
5. Serebryakov V.A. What is a semantic digital library. *Elektronnye biblioteki: perspektivnye metody i tekhnologii, elektronnye kollektsii – RCDL: tr. 16 Vseross. nauch. konf*. [Proc. 16th All-Russian Scientific Conf. on Digital Libraries: Prospective Methods and Technologies, Electronic Collections]. Dubna, Joint Institute for Nuclear Research Publ., 2014, pp. 21–25 (in Russ.).
6. Palano R., Pandurino A., Guido A.L. Conceptual design of web application families: the BWW approach. *Proc. 6th Workshop on Domain Specific Modeling*. Portland, USA. 2006, pp. 23–32.
7. Kruk S.R. *JeromeDL-a Semantic Digital Library*. 2007, 268 p.
8. Miller G. A. WordNet: a lexical database for English. *Communications of the ACM*. 1995, vol. 38, no. 11, pp. 39–41.
9. Kruk S.R., Synak M., Zimmermann K. MarcOnt--Integration ontology for bibliographic description formats. *Intern. Conf. on Dublin Core and Metadata Applications*. 2005, pp. 231–234.
10. *FOAF Vocabulary Specification 0.99*. Available at: http://xmlns.com/foaf/spec/ (accessed June 25, 2016).
11. Isaac A., Summers E. *Skos Simple Knowledge Organization System Primer*. Working Group Note, W3C. 2009.
12. *SPARQL 1.1 Overview*. Available at: https://www.w3.org/TR/sparql11-overview/ (accessed June 25, 2016).
13. Witten I.H., Bainbridge D., Boddie S.J. Greenstone: Open-source digital library software with end-user collection building. *Online Information Review*. 2001, vol. 25, no. 5, pp. 288–298.
14. Aloia N., Concordia C., Meghini C. *Implementing BRICKS, a Digital Library Management System*. SEBD. 2007, pp. 4–15.
15. *Europeana Collections*. Available at: http://www.europeana.eu (accessed June 25, 2016).
16. Doerr M. The europeana data model (edm). *World Library and Information Congr.: 76th IFLA General Conf. and Assembly*. 2010, pp. 10–15.
17. *DBpedia*. Available at: http://dbpedia.org/ (accessed June 25, 2016).
18. Candela L., Castelli D., Dobreva M., Ferro N., Ioannidis Y., Katifori H., Koutrika G., Meghini C., Pagano P., Ross S., Agosti M., Schuldt H., Soergel D. *The DELOS Digital Library Reference Model Foundations for Digital Libraries. IST–2002 2.3.1.12. Technology-enhanced Learning and Access to Cultural Heritage*. Version 0.98, 2007. Available at: http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf (accessed June 25, 2016).
19. Ataeva O.M., Serebryakov V.A. The basic concepts of a semantic libraries formal model and its integration process formalization. *Programmnye produkty i sistemy* [Software & Systems]. 2015, no. 4 (112), pp. 180–187 (in Russ.).
20. Ataeva O.M., Bezdushny A.N. Modeling workflows in a data reduction task. *Informatsionnoe obespechenie nauki. Novye tekhnologii. Sb. nauch. tr.* [Proc. Science Information Support. New technologies]. Kalenov N.E. (Ed.). Moscow, Nauchny Mir Publ., 2009, pp. 323–329 (in Russ.).
21. Castro L.J.G., Giraldo O.X., Castro A.G. Using the Annotation Ontology in semantic digital libraries. *Proc. 9th Int. Semantic Web Conf. (ISWC)*. 2010, pp. 153–156.
22. Kruk S.R., McDaniel B. *Semantic digital libraries*. Heidelberg, Springer Publ., 2009, 245 p.
23. Lee Y., Yang S.Q. *Folksonomies as Subject Access-A Survey of Tagging in Library Online Catalogs and Discovery Layers*. IFLA Pub. Series. Accepted for publication. 2012.
24. Spiteri L.F. The structure and form of folksonomy tags: The road to the public library catalog. *Information Technology and Libraries*. 2013, vol. 26, no. 3, pp. 13–25.