УДК 681.3.06:004.89

А.Ф. Тузовский, Х. Ле

Архитектура электронных библиотек на основе технологий Semantic Web

Описывается новая архитектура электронных библиотек на основе явного описания семантики ресурсов с использованием технологий Semantic Web. Рассматривается способ описания семантики ресурсов на основе единой онтологической модели. Поясняются предлагаемая архитектура семантических электронных библиотек и апробация предложенного подхода в виде программных систем SemArc и SemDL.

Ключевые слова: электронная библиотека, семантические технологии, метаданные, архитектура информационной системы.

Электронная библиотека (ЭБ) – это программная система, которая поддерживает сбор, управление, хранение и распространение электронных ресурсов (документов, мультимедиа данных) для эффективного и долговременного их использования [1]. ЭБ должна описывать большое количество разнородных объектов: электронные ресурсы (ЭР), каталоги, пользователи (читатели, библиотекари), авторы, издательства и т.п. И на основе этих описаний обычно решается набор таких задач, как: формирование метаданных (описаний) всех объектов, автоматическая категоризация ЭР по рубрикам каталога, поиск ЭР по запросам или путем просмотра каталога, навигация между описаниями объектов, формирование и поддержка профилей пользователей, рекомендация новых ЭР читателям.

Решение этих задач в классических ЭБ обычно основывается на реляционных базах данных (БД). Описание всех разнородных объектов выполняется с использованием реляционных схем. В базах данных сохраняются *контекстные данные*, описывающие связи каждого объекта с другими объектами и литералами (например, для документов – название, авторы, издательство, год публикации и т.п.). Кроме этого, в таблицах могут храниться и контентные данные, которые описывают содержание объектов (например, для книги – это код УДК и/или набор ключевых слов, а для читателя ЭБ – это набор терминов, описывающих область его интересов).

Недостатком такого подхода является то, что реляционная схема является жесткой и ее изменение приводит к необходимости изменения большого количества программного кода; на основе реляционных схем невозможно выполнять какие-либо логические выводы; описание контента с помощью кода УДК и набора терминов имеет очень ограниченные возможности для оценки смысловой близости между однотипными и тем более разнотипными объектами. Кроме этого, использование реляционных схем существенно ограничивает взаимодействие между разными ЭБ. В связи с этим требуется разработка новых архитектур ЭБ.

Технологии Semantic Web. Для решения перечисленных проблем и повышения эффективности работы электронных библиотек становится необходимо описывать семантику (смысл) содержания ЭР и реализовывать работу с ними путём использования моделей представления знаний, например, таких, как семантические сети, фреймы и онтологии. В настоящее время более совершенными и выразительными считаются онтологические модели. Это объясняется тем, что они основываются на формальных (дескрипторных) логиках, и для работы с ними в рамках концепции Semantic Web [2] разработан набор стандартных технологий, позволяющих создавать, описывать и использовать онтологические модели. К технологиям Semantic Web относятся такие языки, как RDF (Resource Description Framework), RDFS (RDF Schema), OWL (Ontology Web Language) и SPARQL (Simple Protocol And RDF Query Language). Кроме этого, разработано большое количество программных систем по работе RDF-данными и онтологическими моделями (редакторы, системы логического вывода, RDF-хранилища и т.п.).

Язык RDF задает формат описания данных в виде простых утверждений типа (субъект, предикат, объект), называемых триплетами. Все элементы триплетов могут задаваться с помощью URI-идентификаторов. Смысл элементов (их разделение на классы и взаимосвязи между собой) задается с помощью онтологической модели, которая может описывать на таких языках, как RDFS и OWL [3]. При этом сами языки RDFS и OWL основываются на языке RDF, и их описания являются набо-

рами триплетов. Таким образом, и метаданные (контентные и контекстные), и логические модели, на которых они основываются, являются триплетами и могут храниться в RDF-хранилищах.

В настоящее время разработаны RDF-хранилища, позволяющие содержать и использовать миллиарды триплетов [4]. Кроме этого, разработано большое количество программных систем по работе с онтологическими моделями (редакторы, системы логического вывода и т.п.).

Семантические электронные библиотеки. На основе результатов исследований, выполненных в области ЭБ, технологий Semantic Web, социальных сетей и организации взаимодействия человека с компьютером, разрабатываются *семантические электронные библиотеки.* В данном направлении уже выполнялось большое количество исследований, например: GREENSTONE [5] и DELOS [6]. Однако в этих системах технологии Semantic Web в основном использовались для более точного описания контекста ЭР, а их контент описывается с помощью понятий онтологий без достаточно подробного пояснения содержания ресурсов. Реализация функций электронных библиотек с использованием таких описаний приводит к недостаточной точности результатов их работы.

Можно выделить следующие уровни описания контента электронных ресурсов: 1) с помощью набора терминов; 2) с помощью набора понятий (текстовых меток понятий онтологий), связанных между собой на основе единой онтологии; 3) с помощью набора утверждений (триплетов), основанных на некоторой единой онтологии. В данной статье предлагается архитектура семантических ЭБ, использующих третий уровень описания контента ЭР, которая основывается на следующих положениях:

1. Описание всех информационных ресурсов ЭБ должно основываться на использовании единой онтологической модели.

2. Содержание (контент) информационных ресурсов ЭБ описывается метаданными, состоящими из простых утверждений (триплетов), связанных с единой онтологической моделью для более точного определения используемых в документах понятий и связей между ними.

3. Основные функции ЭБ (поиск, категоризация, формирование рекомендаций и др.) реализуются на основе семантических контекстных и контентных метаданных.

Онтологии семантической электронной библиотеки. Основной идеей исследуемого подхода является использование онтологий предметных областей для аннотирования содержания электронных ресурсов [7, 8]. Под онтологией O понимается знаковая система $\langle C, P, I, L, T\rangle$, где C – множество элементов, которые называются понятиями; P – множество элементов, называемых свойствами (двуместными предикатами); I – множество экземпляров понятий; L – множество текстовых меток или значений понятий и свойств; T – частичный порядок на множестве C и P. С помощью набора предикатов P онтологии могут описываться различные отношения между различными понятиями и экземплярами. Такие отношения задаются с использованием простых утверждений $\langle s, p, o\rangle$, где s и o – это субъект и объект (экземпляр) высказывания, а $p \in P$ – это предикат. Субъекты и предикаты описываются в онтологии O. Считаем, что любому свойству $p \in P$ может быть задан весовой коэффициент (семантический вес) $pv \in [0, 1]$, задающий смысловую близость между субъектом и объектом утверждения (1 – субъект и объект считаются похожими по семантике, а 0 – не связанными), составленного с использованием данного свойства.

Описание всех информационных ресурсов ЭБ основывается на единой онтологии $O = \{O_б, O_п\}$. К базовым онтологиям $O_б$ относятся: *онтология пользователей*, *онтология ресурсов* и *онтология системы*. Подробное их описание приведено в [7]. К онтологиям $O_п = \{O_1, \ldots, O_m\}$ относится иерархически организованная, последовательно расширяемая система онтологий основных областей знаний для описания содержания ЭР. Они используются для аннотирования содержания ЭР. Содержание ЭР относится к конкретным предметным областям знаний, которые должны быть достаточно точно описаны.

Семантические модели информационных ресурсов. Любой объект в ЭБ описывается метаданными, набор которых представляет собой специальным образом структурированную информацию, характеризующую их содержание. Метаданные используются для создания метаописаний – это определенный набор метаданных, поставленных в соответствие какому-либо ресурсу, и синтаксических правил, определяющих способ представления метаданных в метаописании: правило их оформления, порядок следования и т.п.

Каждый ресурс может быть рассмотрен с двух точек зрения: контекста и контента. *Семантическими метаданными* ресурса называется его описание относительно некоторой формальной модели О, определяющей семантику рассматриваемой области знаний. На основе используемой онтологи-

ческой модели знаний O описание любого ресурса $s \in I$ может быть представлено в виде следующего набора семантических метаданных: $M(s) = (M_к(s), Mc(s))$, где $M_к(s)$ – это *контекстные метаданные* ресурса знаний s, описывающие его взаимосвязи с другими ресурсами, понятиями системы или литералами, а $M_с(s)$ – *контентные метаданные* ресурса знаний s, описывающие знания и информацию, содержащиеся в ресурсе s.

Контекстные метаданные ресурса s (заданного URI-идентификатором) – это набор простых утверждений (триплетов) $M_к = \{t_i = \langle s, p_i, o_i \rangle \mid i \in [1, n]\}$, где $s \in I$ – это аннотируемый ресурс (субъект), $o \in C \cup I \cup L$ – объект утверждения, $p \in P$ – отношение между субъектом и объектом. *Контентные метаданные* ресурса – это набор простых утверждений (кортежей) $M_с = \{t_j = \langle s_j, p_j, o_j, v_j \rangle \mid j \in [1, m]\}$, где $s \in C \cup I$ – это субъект утверждения, $o \in C \cup I$ – объект утверждения, $p \in P$ – отношение между субъектом и объектом, а v – весовой коэффициент, который оценивает значимость данного утверждения.

Онтологические и семантические модели ЭБ хранятся в базе знаний (БЗ). Под БЗ понимается множество триплетов, хранящихся в RDF-хранилище, которые описывают онтологии, метаописания и экземпляры понятий. БЗ включает: онтологические модели (O); метаописания (M); экземпляры понятий (I). Экземпляры понятий описываются наборами свойств: свойства типов данных (DatatypeProperty); объектные свойства (ObjectProperty) – связи между экземплярами.

Учитывая, что модели описаний ресурсов связаны за счет использования единой онтологической модели ресурсов ЭБ, имеется возможность оценки их подобия (сходства) между собой на основе некоторой метрики семантического подобия $\mathrm{Sim}(M_i, M_j)$, с использованием которой могут быть разработаны различные методы решения основных задач ЭБ.

Семантические ЭБ используют следующие виды метаописаний:

- *Метаописания электронных ресурсов* (документы, мультимедийные объекты), состоящих из контекстных метаданных (набора триплетов, описывающих контекст) и контентных метаданных (наборов триплетов, описывающих контент). Создание метаописаний мультимедийных объектов выполняется только ручным способом, а создание метаописаний текстовых документов – с помощью полуавтоматического способа.
- *Метаописания разделов каталогов* (рубрик). Метаописание рубрики также описывается контекстными и контентными метаданными из всех ресурсов, входящих в шаблон описания данной рубрики.
- *Метаописания профилей пользователей*: набор триплетов, описывающих информацию, связанную с пользователем, и интересы. Интересами пользователя в ЭБ могут быть конкретные электронные ресурсы, их авторы, издания и т.п.
- *Описания поисковых запросов*: в семантической ЭБ запросы описываются в виде наборов триплетов, описывающих как контекст, так и контент искомого ресурса.

Таким образом, все ресурсы описываются с помощью метаописаний одинаковой структуры на основе единой онтологической модели семантической ЭБ. Исходя из этого, можно выполнять оценку семантических близостей между ними. На основе определения семантической близости между метаописаниями разработан набор методов решения основных задач семантической ЭБ. Подробное описание этих методов можно найти в [8, 14].

Архитектура семантических электронных библиотек. На основе анализа архитектур существующих реализаций ЭБ и обобщенной архитектуры информационной системы, основанной на технологиях Semantic Web, разработана многоуровневая обобщенная архитектура семантической ЭБ, которая показана на рис. 1.

Предложенная архитектура разделена на 6 уровней, соответствующих различным этапам обработки запросов пользователей системы. *Уровень представления системы* поддерживает взаимодействие с web-браузерами клиентов и мобильными клиентскими приложениями, выполняющими запросы к системе. *Уровень представления данных* подготавливает клиентам данные в различных форматах, например таких, как HTML/AJAX для веб-браузеров, RDF/XML – для других приложений. *Уровень подготовки данных* содержит средства проверки соответствия данных требованиям сервисов, а также правам доступа к ним конечных пользователей.

Основной частью системы является *уровень бизнес-логики*, в котором реализуются все функции (сервисы) системы. Все функции (сервисы) семантической ЭБ разделены на следующие категории: сервисы, доступные обычным пользователям системы (читателям): *навигация*, *рекомендация*, *кате-*

горизация и *семантический поиск*, *профилирование пользователей*; сервисы, доступные привилегированным пользователям (библиотекарям) системы: *аннотирование*, *администрирование*, *ведение онтологической базы знаний*; *вспомогательные сервисы*; сервисы, предназначенные для поддержки работы других сервисов: *индексирование*, *вычисление семантических близостей*, *ранжирование*, *ведение онтологической базы знаний*. *Уровень абстрактных моделей объектов* включает описания всех объектов, процессов, провайдеров к разным источникам данных или сервисов системы с помощью набора интерфейса и классов. *Уровень источников данных* семантической ЭБ включает различные RDF-хранилища и базы индексов документов, URI-идентификаторов объектов знаний и предварительные оценки их семантической близости.

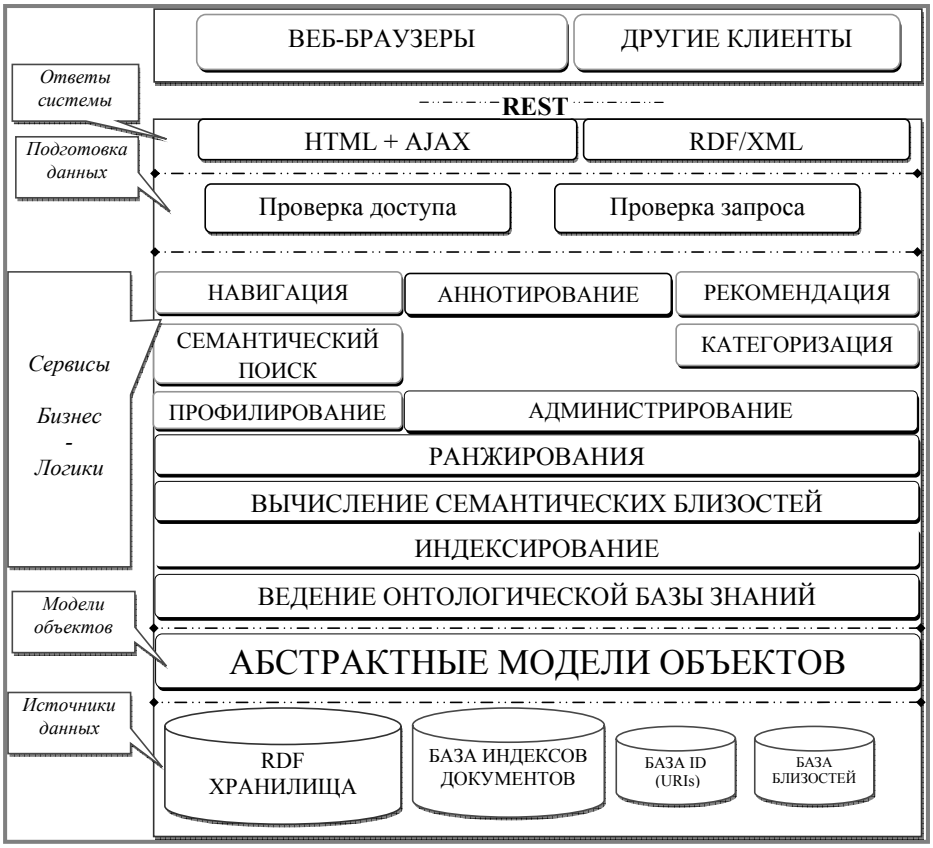

Рис. 1. Многоуровневая архитектура семантической ЭБ

Апробация предлагаемой архитектуры семантических электронных библиотек. Для апробации предлагаемого подхода, созданных моделей и методов были разработаны две программные системы: SemARC – система семантического аннотирования ресурсов и SemDL – система управления хранилищем электронных ресурсов с использованием семантических технологий. На эти программные системы были получены свидетельства о регистрации программы для ЭВМ в Роспатенте РФ №2013613146 и 2013613266.

Система семантического аннотирования ресурсов SemARC предназначена для семантического аннотирования ЭР наборами триплетов. Она выполняет такие функции, как визуализация и навигация между компонентами триплетов проаннотированных ЭР; автоматический анализ текстов в форматах *txt, doc, pdf, xml, html* с возможностью выделения именных сущностей и их сопоставления с экземплярами онтологий, хранящихся в БЗ; автоматическое и ручное разрешение семантической многозначности; формирование триплетов путем выбора их отдельных компонентов из онтологической базы знаний с помощью поисковых запросов. Обобщенная структура системы SemArc показана на рис. 2.

Программная система SemArc создана на основе технологий Java Server Pages и Java Servlet и может использоваться либо автономно, либо как HTML-тег на странице JSP для аннотирования информационных объектов ЭБ. В качестве системы управления базой знаний, реализованной с помощью RDF-хранилища, использовалась система Sesame [10]. Данная система представляет собой

web-сервис с открытым исходным кодом на языке Java для хранения триплетов, описывающих все понятия и экземпляры онтологий библиотеки.

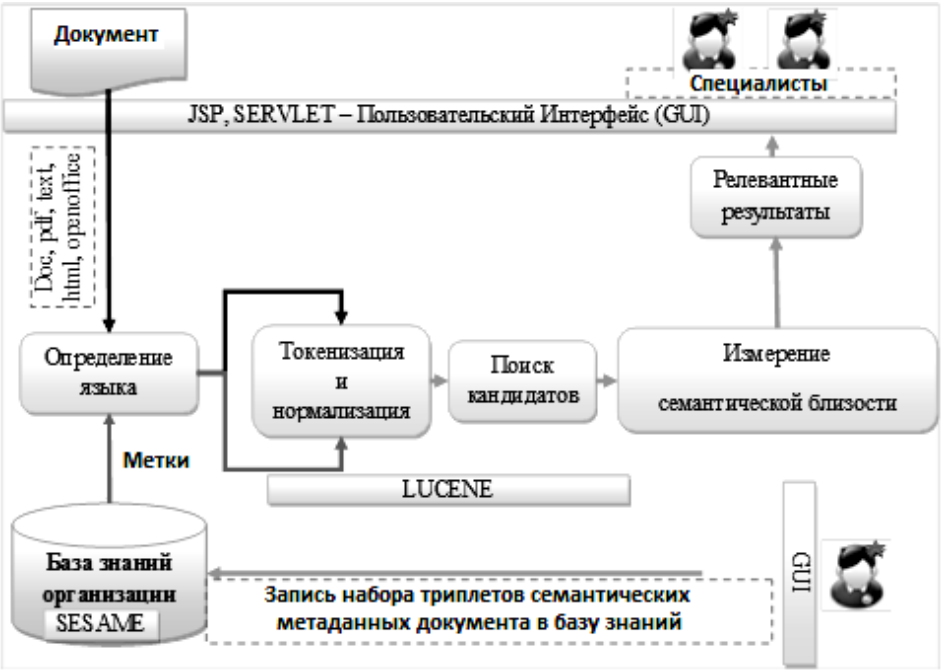

Рис. 2. Программная архитектура системы SemARC

Для определения языка, используемого в документе, и меток каждого понятия или экземпляра используется пакет LangDetect [11]. Данный пакет является библиотекой с открытым исходным кодом на языке Java, которая позволяет идентифицировать большое количество естественных языков, в том числе и русский. Результат работы данного пакета используется для вызова конкретных компонентов анализа текста. Наиболее важной частью системы SemARC является компонент Lucene [12] – библиотека полнотекстового поиска на языке Java. Она позволяет выполнить индексацию документов различных форматов (с помощью специальных модулей).

Система управления хранилищем электронных ресурсов с использованием семантических технологий SemDL предназначена для управления хранилищем электронных ресурсов (документы, изображения, аудио- и видеофайлы) на основе использования таких семантических технологий, как RDF, RDFS, OWL и SPARQL. Программа обеспечивает выполнение регистрации пользователей и создания их профилей на основе анализа интересов к ЭР; ведение каталогов, содержащих аннотации ЭР в виде набора триплетов; просмотр электронных ресурсов и выполнение переходов между информационными объектами (документами, профилями пользователей, категориями понятий и т.д.); категоризации ЭР на основе их семантических аннотаций; формирование рекомендаций электронных ресурсов на основе интересов пользователя и семантических описаний ресурсов; выполнение семантического поиска электронных ресурсов. Архитектура системы SemDL показана на рис. 3.

Как видно из рис. 3, система SemDL разделена на следующие уровни: *Web-интерфейс*, *Контроль сеанса и права доступа к функциям системы*, *Вызов функций*, *Компоненты и пакеты*, *Источники данных*. Пользователи взаимодействуют с системой с помощью web-интерфейса и могут вызывать доступные им функции.

Ведение онтологической базы знаний осуществляется с помощью приложений с web-интерфейсом, которые позволяют управлять базами знаний (RDF-хранилищами) на сервере Sesame. Исходные RDF-данные могут создаваться с помощью редактора онтологий Protégé [13].

Поиск в системе SemDL реализуется разными способами – простой полнотекстовый и более сложный семантический с ранжированием результатов. В ходе формирования запросов показываются понятия онтологии, которые помогают пользователям лучше их сформулировать. Семантический поиск в SemDL выполняется отдельно по контексту или контенту ресурса, при этом входным данным является набор триплетов запроса. Результаты поиска ранжируются в порядке убывания оценки их релевантности.

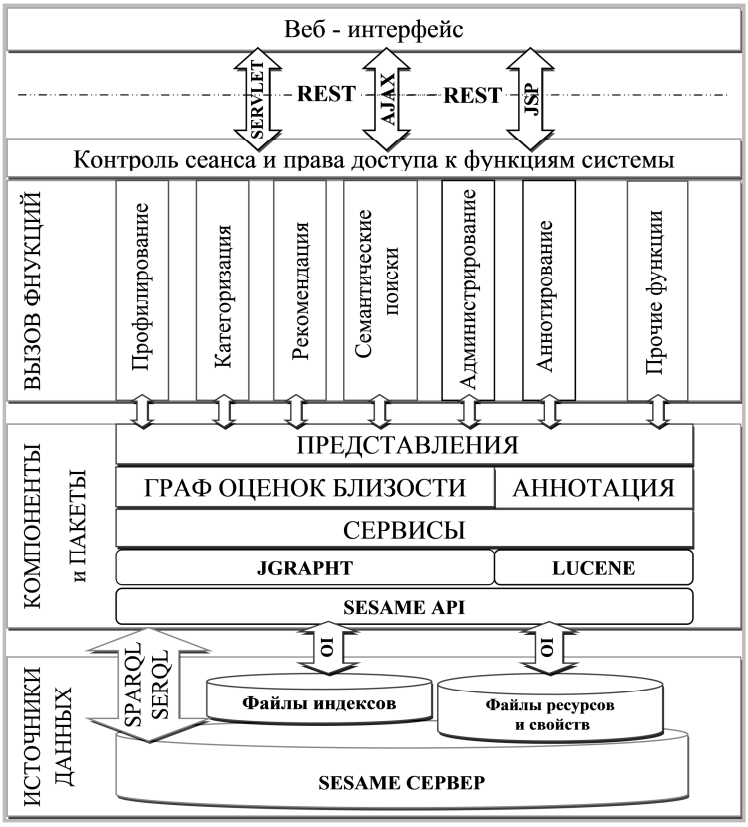

Рис. 3. Программная архитектура системы SemDL

Заключение. Развитие электронных библиотек требует создания подходов к работе с явным описанием содержания ЭР. В данной статье предложен такой подход на основе использования технологий Semantic Web. Реализация предложенного подхода в виде программных систем SemArc и SemDL и результаты ее тестирования [14] показали реализуемость данного подхода и эффективность решения таких основных задач ЭБ, как поиск, категоризация и рекомендация.

Статья подготовлена в ходе выполнения госзадания «Наука», код проекта 2551 «Развитие архитектуры информационных систем на основе технологий Semantic Web».

Литература

1. Электронные библиотечные системы [Электронный ресурс]. – 2014. – Режим доступа: http://library.mstu.edu.ru/resources/files/ebs.pptx, свободный (дата обращения: 01.07.2014).

2. Antoniou G. A Semantic Web Primer / G. Antoniou, F. Harmelen. – London: The MIT Press Cambridge, 2008. – 287 p.

3. Allemang D. Semantic Web for the Working Ontologist (Modeling in RDF, RDFS and OWL) / D. Allemang, J. Hendler. – New York: Morgan Kaufmann Publishers, 2011. – 364 p.

4. Large Triple Stores [Электронный ресурс]. – 2014. – Режим доступа: http://www.w3.org/wiki/LargeTripleStores, свободный (дата обращения: 01.07.2014).

5. About Greenstone [Электронный ресурс]. – 2014. – Режим доступа: http://www.greenstone.org/, свободный (дата обращения: 01.07.2014).

6. Welcome to the DELOS Network of Excellence [Электронный ресурс]. – 2014. – Режим доступа: http://delos.info/, свободный (дата обращения: 01.07.2014).

7. Ле Х. Использование онтологий в электронных библиотеках / Х. Ле, А.Ф. Тузовский // Изв. Том. политех. ун-та. – 2012. – Т. 320. – № 5. – С. 36–42.

8. Ле Х. Разработка семантических электронных библиотек на основе онтологических моделей / Хоай Ле, А.Ф. Тузовский // Сб. докл. XV Всерос. науч. конф. RCDL'2013. – Ярославль, 2013. – С. 291–299.

9. Tran T. Lifecycle-Support in Architectures for Ontology-Based Information Systems / T. Tran, P. Haase, H. Lewen et al. // The Semantic Web. Lecture Notes in Computer Science. – 2007 – Vol. 4825. –

10. OpenRDF Sesame Core 2.6.10 API [Электронный ресурс]. – Режим доступа: http://openrdf.callimachus.net/sesame/2.6/apidocs/index.html, свободный (дата обращения: 01.07.2014).

11. Language-detection [Электронный ресурс]. – Режим доступа: http://code.google.com/p/language-detection/, свободный (дата обращения: 01.07.2014).

12. Paul T. The Lucene Search Engine [Электронный ресурс]. – 2004. – Режим доступа: http://www.javaranch.com/journal/2004/04/Lucene.html, свободный (дата обращения: 01.07.2014).

13. A free, open-source ontology editor and framework for building intelligent systems [Электронный ресурс]. – Режим доступа: http://protege.stanford.edu/, свободный (дата обращения: 01.07.2014).

14. Ле Х. Исследование и разработка электронных библиотек на основе явного описания семантики ресурсов с использованием технологий Semantic Web: дис … канд. техн. наук. – Новосибирск, 2014. – 182 с.

Тузовский Анатолий Федорович
Д-р техн. наук, проф. каф. оптимизации систем управления НИТПУ
Тел.: +7 913-822-1952
Эл. почта: tuzovskyaf@tpu.ru

Ле Хоай
Канд. техн. наук, исследователь Центра высокотехнологической и технической коммуникации,
Ханой, Вьетнам
Тел.: +84-969-105-986
Эл. почта: lehotomsk@yahoo.com

Tuzovsky A.F., Le H.
Architecture of digital libraries based on Semantic Web technologies

The paper describes a new architecture of digital libraries based on explicit content declaration using Semantic Web technologies. The method of resource semantic description based on whole ontology is considered. Testing of suggested architecture, the method and models are explained by SemArc and SemDL software systems.
Keywords: digital libraries, semantic technologies, metadata, architecture of information system.