УДК 681.3.06: 004.89

Ле Хоай, А.Ф. Тузовский

Разработка семантических электронных библиотек

Рассматривается подход к созданию электронных библиотек (ЭБ) и их разработки с использованием семантических технологий. Поясняются функции электронных библиотек, для автоматизации которых требуется использовать семантику и предлагается архитектура семантической электронной библиотеки (СЭБ), а также описываются основные выполняемые в ней, процессы.

Ключевые слова: электронная библиотека, семантические технологии, архитектура электронной библиотеки.

В настоящее время имеется много разнородных электронных документов, доступных в компьютерных сетях. В связи с этим становится актуальной проблема организации работы с такими документами и содержащейся в них информацией, используя современные информационные технологии. Необходимо выполнять работу не столько с файлами, в которых содержатся документы, но и с их смыслом, содержащейся в них семантикой.

Под электронными библиотеками понимаются информационные системы, которые автоматизируют решение основных проблем организации работы с документами. Уже достаточно давно предпринимались попытки разработки подходов к созданию электронных библиотек. Однако в связи с тем, что появляются новые требования и новые возможности, связанные с появлением новых технологий, необходимо разрабатывать и новые подходы к созданию ЭБ.

С появлением семантических технологий, позволяющих выполнять работу с семантикой документов, возникла возможность разработки новых подходов к автоматизации работы с электронными документами на новом уровне. В данной статье рассматриваются проблемы функционирования электронных библиотек, на основе которых разработана их модель, основанная на использовании семантических технологий. Также предлагается упрощенная архитектура ЭБ с использованием семантических технологий и описываются основные выполняемые в ней процессы.

Проблема электронных библиотек. В соответствии с [1] можно сделать следующее определение ЭБ. ***Электронные библиотеки*** это организации, в том числе специализированный персонал, представляющие доступ читателей к электронным ресурсам. Кроме того они выполняют отбор, структурирование, предоставление интеллектуального доступа, интерпретацию, распространение, сохранение целостности и обеспечение сохранности в течение длительного времени наборов электронных документов для удобного доступа к ним определенным сообществам специалистов.

В соответствии с данным определением основными компонентами ЭБ являются: специалисты, информационные ресурсы (документы) и информационные технологии.

Электронные библиотеки реализуют набор функций для обеспечения читателям полного доступа к множеству распределенных и разнородных документов, содержащих информацию и знания, интегрируя их в единое информационное пространство [2].

В [2–4] описаны некоторые проблемы ЭБ, основными из которых являются следующие:

- Проблема интеграции разнородной информации (электронных ресурсов, пользовательских профилей, таксономий) на основе различных метаданных, содержащих выразительные семантические описания.
- Проблема поддержки взаимодействия с другими информационными системами (и не только ЭБ) либо с помощью метаданных, либо на уровне коммуникации или с помощью обеих возможностей. При этом в качестве единого языка взаимодействия между системами может использоваться язык RDF (Resource Description Framework).
- Проблема обеспечения надежного, удобного и адаптируемого поиска и интерфейсов просмотра электронных документов, усиленных работой с семантикой.

Для решения таких проблем и улучшения функционирования ЭБ требуется разработать новый тип ЭБ на основе использования новых информационных технологий, в том числе семантических технологий. В этом случае такие ЭБ можно называть семантическими электронными библиотеками.

Существующие подходы. Семантическая электронная библиотека является следующим поколением ЭБ, которые могут быть определены следующим образом [5]: семантическими называются электронные библиотеки, которые созданы на основе результатов исследований, проведенных в области ЭБ, семантических сетей, социальных сетей и организации взаимодействия человека с компьютером. Они объединяют системы организации знания ЭБ с семантическими технологиями и социальными сетями (Web 2.0).

Семантические технологии позволяют поддерживать точность аннотирования электронных документов и возможность интероперабельности различных сервисов. Подход Web 2.0 дает пользователям возможность участвовать в аннотировании и процессе обмена знаниями, что повышает полезность СЭБ (рис. 1).

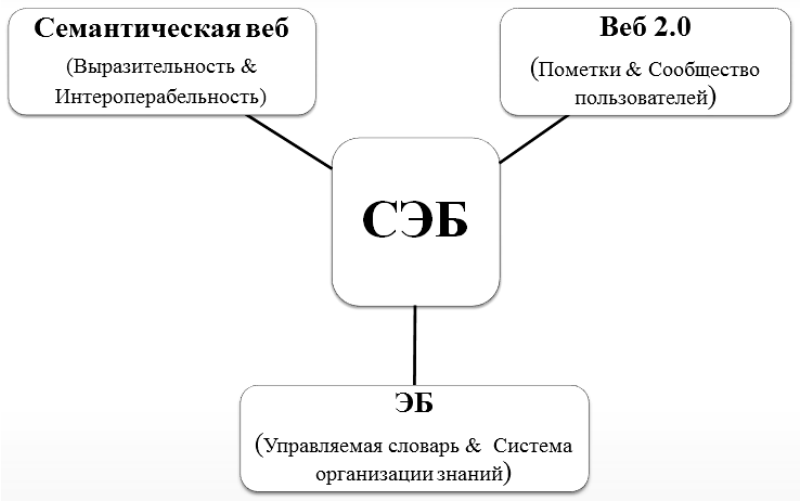

Рис. 1. Семантические электронные библиотеки

В настоящее время ведутся активные исследования в данной области. Примерами проектов, в которых исследуется использование семантических технологий для создания и функционирования ЭБ, являются:

- Проект **SIMILE:** расширяет и повышает возможности такой известной системы организации работы с электронными ресурсами, как DSpace [6], что позволяет улучшить интероперабельность между хранилищами электронных ресурсов, между схемами-словарями-онтологиями, повысить качество метаданных и услуг. Ключевая проблема заключается в том, что совместно используемые коллекции документов, взаимодействуют между посредством личностей, сообществ и хранилищ организаций. В проекте SIMILE предлагается подход к описанию таких взаимодействий с использованием семантических технологий.
- Проект **GREENSTONE:** программное обеспечение с открытыми кодами для создания и распространения электронных библиотечных коллекций. **GREENSTONE** предоставляет динамическую информационную систему управления ЭБ, которая может гибко настраиваться во время выполнения. Это позволяет уменьшить накладные расходы, связанные с созданием коллекции.
- Проект **DELOS**: является сетью передового опыта в области ЭБ и частично финансируется Европейским сообществом в рамках программы разработки информационных технологий. Основными задачами являются проведение и обмен результатами исследований по созданию технологий разработки следующих поколений ЭБ, на основе таких технологии, как, например, P2P, Grid, SOA.
- Проект **BRICKS:** нацелен на создание организационных и технологических основ сети ЭБ, которая позволит выполнять обмен знаниями и ресурсами в области культурного наследия. Для описания метаданных используется язык RDF, для описания схем используется язык OWL, а запросы к метаданным выполняются с использованием языка SPARQL.
- Проект **JEROMEDL:** представляет собой социальную семантическую библиотеку, которая использует технологии Semantic Web и социальных сетей для повышения как интероперабельности, так и удобства работы. Данный проект нацелен на решение таких проблем, как интеграция и поиск информации из различных библиографических источников, а также на решение проблемы связывания знаний различных пользователей.

Задачи разработки электронных библиотек с использованием семантических технологий. На основе анализа выполненных проектов и известных публикаций можно сформулировать следующие задачи построения ЭБ с использованием семантических технологий:

1) разработка базовой архитектуры ЭБ с использованием семантических технологий;
2) разработка методов сбора и объединения онтологий (RDFS и OWL) для каталогов электронной библиотеки;
3) разработка методов формирования и объединения метаданных (RDF) документов, содержащихся в ЭБ;
4) разработка методов ведения базы знания ЭБ с метаданными и онтологиями;
5) разработка методов описания поисковых запросов пользователей к каталогу ЭБ;
6) разработка методов оценки семантической близости описаний запросов пользователей и метаданных документов библиотеки;
7) разработка методов поиска метаданных наиболее соответствующих составленному пользователем запросу;
8) разработка программной системы, реализующей все описанные выше процессы.

При исследовании и решении данной задачи необходимо учитывать контроль доступа пользователей различных уровней к документам и операциям с ними. Дальше, при разработке онтологий, как можно больше использовать уже разработанные онтологии. Кроме этого, должна быть возможность информировать пользователей о поступлении новых документов. В ходе работы могут появиться и новые требования, которые должны быть решены.

Предлагаемая архитектура и основные процессы. Базовая архитектурная модель для любой ЭБ состоит из трех основных элементов [5]:

- классы пользователей;
- уровни данных и сервисов;
- используемые информационные технологии.

Классы пользователей: семантические электронные библиотеки, как системы, управляют различными ресурсами и организуют различные способы контроля доступа к ним. Существуют некоторые работы, которые выполняют только специальные пользователи, такие как библиотекари и администраторы библиотеки или пользователи с высоким рейтингом работы с библиотекой.

Таким образом, семантические электронные библиотеки должны предоставлять возможность для работы различным класс пользователей, которые существуют и в обычных библиотеках.

Уровни данных и сервисов: архитектура семантических электронных библиотек может быть разделена на шесть уровней:

1) пользовательский уровень;
2) уровень представления данных;
3) уровень подготовки данных;
4) уровень работы с данными;
5) уровень абстрактных описаний данных;
6) уровень доступа к источникам данных.

Используемые информационные технологии: в данный компонент могут быть включены такие технологии, как: SOA (архитектуры, ориентированные на сервисы), P2P (одноранговые сети) и Grid (коллективное использование вычислительных ресурсов). SOA облегчает выполнение задач бизнес-логики. P2P должно быть включено для обеспечения операционной совместимости между распределенными семантическими электронными библиотеками. Гриды поддерживают коллективное выполнение некоторых услуг на распределенных вычислительных ресурсах для повышения эффективности работы библиотеки.

На основе сделанного описания задач модели ЭБ с использованием семантических технологий, может быть предложена следующая ее архитектура (рис. 2), в которой представлены пользователи, интерфейсы, две функции и физический уровень данных – файловое хранилище и база знаний.

В данной архитектуре выполняются два основных процесса:

Процесс добавления объектов: сначала пользователи с соответствующими правами доступа через интерфейс ЭБ вызывают функцию добавления объектов. После этого, пользователям предоставляется соответственный интерфейс для заполнения данных об информационном объекте (документ, видео, аудио, событие…).

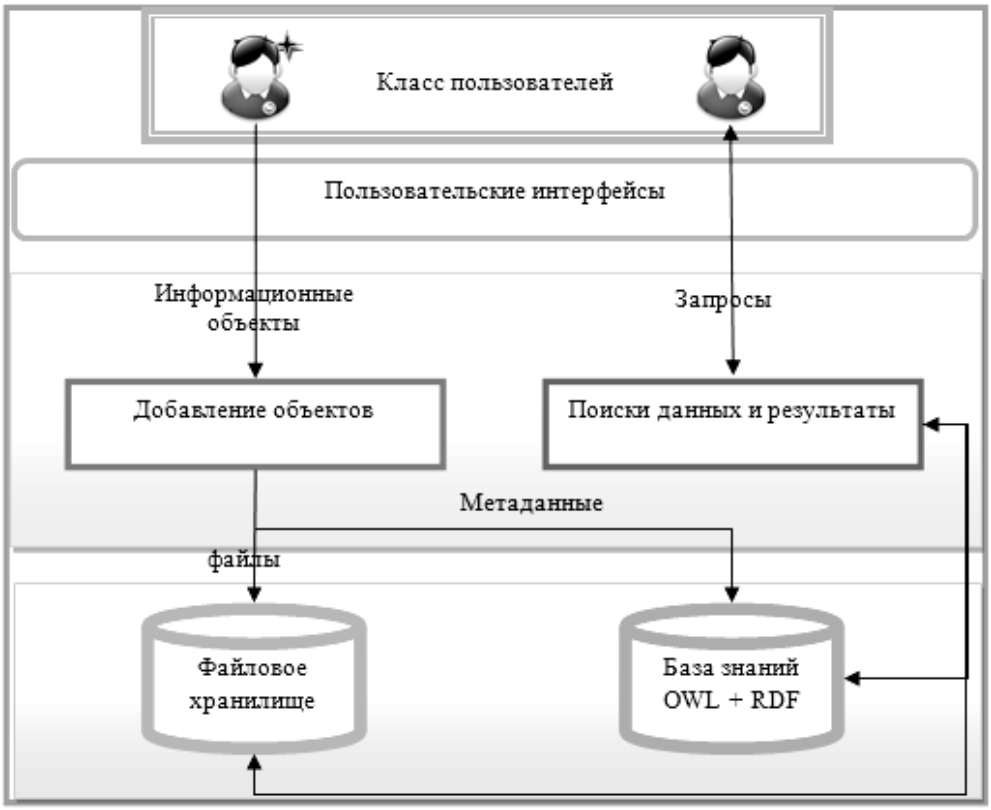

Рис. 2. Предложенная архитектура

Для описания метаданных необходимы некоторые данные в соответствии с используемой онтологией, например: автор, заголовок, ключевые слова, аннотация и др. Затем эти метаданные передаются в базу знаний, где хранятся в формате **RDF** или **OWL**, а сами документы сохраняются в файловом хранилище.

В этот процесс можно включить систему организации знаний [7], которая представляет управляемые словари для аннотирования ресурсов и их описания. Одной особенностью научных статьей является аннотация, которая представляет собой некоторые текстовые строки об этих статьях, и проблема видится в том, как можно извлекать смыслы этого текста. Одним из подходов является использование GATE, как в [8]. GATE позволяет превратить тексты на естественном языке в описания на языке RDF, и затем эти данные передаются в базу знаний.

Процесс поиска данных и возврата результатов: в этом процессе любые пользователи ЭБ или только зарегистрированные могут использовать функцию поиска. Поиск данных осуществляется с помощью пользовательского интерфейса.

Пользователи вначале могут выбрать поля для осуществления поиска в зависимости от данных, которые были заложены в базу знаний. Функция, выполняющая поиски данных, запрашивает доступ к базе знаний и файловому хранилищу и отправляет запрос (на языке **SPARQL**).

После этого запрошенные данные возвращаются, сортируются и показываются пользователю с использованием графического интерфейса. ЭБ может предоставить поиск на естественном языке, при этом пользовательские запросы будут проанализированы с помощью особенной функции для трансляции их в **SPARQL**-запросы [9].

Существуют и другие процессы, которые осуществляются в ЭБ. Такими процессами могут быть службы оповещения, которые информируют пользователей о поступлении новых или изменении существующих информационных объектов в соответствии с их интересами. При необходимости некоторые функции для ЭБ могут быть расширены.

Заключение. В данной статье были описаны начальные шаги по задаче построения модели ЭБ с использованием семантических технологий. В дальнейшем предполагается разработать формат метаданных для создания каталогов документов и самых документов. Основной задачей создания

СЭБ будет организация семантического поиска и определения соответствия метаданных составленному запросу поиска.

Литература

1. Shiri A.A. Digital library research: current developments and trends // Library Review. – 2003. – Vol. 52. – P. 198–202.

2. McDaniel Bill. Semantic Digital Library / Bill McDaniel, Sebastian Ryszard Kruk. – Springer, 2009. – P. 79–80.

3. Ding Hao. A semantic search framework in peer-to-peer based digital libraries. – NTNU, Norway, 2006. – P. 106–108.

4. Sukhdev Singh. Digital Library: Definition to Implementation [Электронный ресурс]. – Режим доступа: http://arizona.openrepository.com/arizona/bitstream/10150/106534/1/lecture_rcc_26jul03.pdf, свободный (дата обращения: 20.04.2011).

5. Semantic digital library web. [Электронный ресурс]. – Режим доступа: http://semdl.info/, свободный (дата обращения: 20.04.2011).

6. WikiPedia. Dspace. [Электронный ресурс]. – Режим доступа: http://en.wikipedia.org/wiki/DSpace, свободный (дата обращения: 20.04.2011).

7. McDaniel Bill. Semantic Digital Library / Bill McDaniel, Sebastian Ryszard Kruk. – Springer, 2009. – P. 23–26.

8. Ontology-Driven Semantic Digital Library / Shahrul Azman Noah, Nor Afni Raziah Alias, Nurul Aida Osman et al. // Lecture Notes in Computer Science. – 2010. – Vol. 6458/2010. – P. 141–150.

9. Linckels Serge. E-Librarian Service: User-Friendly Semantic Search in Digital Libraries / Serge Linckels, Christoph Meinel. – Springer, 2011. – P. 118–119.

Ле Хоай
Аспирант каф. оптимизации систем управления
Национального исследовательского Томского политехнического университета (НИТПУ)
Тел.: 8-913-108-01-44
Эл. почта: lehotomsk@yahoo.com

Тузовский Анатолий Федорович
Д-р техн. наук, профессор каф. оптимизации систем управления НИТПУ
Тел.: +7 (382-2) 42-14-85
Эл. почта: tuzovskyaf@tpu.ru

Hoai Le, Tuzovsky A.F.
Development of semantic digital libraries

This article discusses an approach to the creation of digital libraries (DL) and development using semantic technologies. We explain the functions of digital libraries for automation of which we propose the architecture and the semantics of Semantic Digital Library (SEB), also the basic processes are described.
Keywords: digital library, semantic technologies, the architecture of the electronic library.