Подход к созданию персональной семантической электронной библиотеки

© О. М. Атаева © В. А. Серебряков
Вычислительный центр им. А.А. Дородницына РАН,
Москва
oli@ultimeta.ru serebr@ultimeta.ru

Аннотация

Целью данной работы является разработка информационной системы для создания семантической электронной библиотеки, наполнение которой индивидуально для каждого пользователя системы и выполняется из разнородных источников данных, расположенных на просторах сети и интегрированных в облако LOD. В работе представлена общая схема системы, выделены ее основные модули и дана краткая характеристика каждого из них. Основная задача системы заключается в предоставлении пользователю унифицированного представления для возможности извлечения интересующей пользователя информации по определенной предметной области. Эта предметная область описывается в терминах тезауруса, поддержка которого обеспечивается соответствующим модулем. Обсуждается задача отображения терминов источников данных на этот тезаурус.

1 Введение

Обилие информационных источников в сети вызывает трудности при поиске ресурсов. Поиск ресурсов по определенной тематике отдельно в каждом источнике требует времени, не менее затратно использовать для этого обычные поисковые системы, которые в результатах поиска выдают много «мусора». В связи с этим уже давно возникла насущная необходимость в структурировании информации в сети. Последнее десятилетие наблюдается бурное развитие технологий Semantic Web и активное развитие сообщества, поддерживающего идею Linked Open Data (LOD). Эти события оказали влияние и на электронные библиотеки, которые трансформируются и превращаются центры данных, вокруг которых формируется сообщество заинтересованных экспертов и пользователей, принимающих активное участие в жизни таких динамически развивающихся библиотек. Основной задачей таких центров является интеграция контента различных электронных библиотек, что позволяет увеличить степень повторного использования данных, понизить степень дублирования данных, повысить ценность данных за счет связывания их с другими данными..

Целью данной работы является разработка информационной системы для создания семантической электронной библиотеки, наполнение которой индивидуально для каждого пользователя системы и выполняется из разнородных источников данных, расположенных на просторах сети и интегрированных в облако LOD. Наполнение происходит полуавтоматически, при этом пользователь может не быть осведомлен о структуре данных источника. Система должна быть «проста» в использовании, т.е. не требовать от пользователя специальных знаний. Тематика поисковых запросов по пространству LOD определяется пользователем, с использованием внешних источников данных, в качестве которых, например, могут выступать другие библиотеки. Из результатов поиска пользователь может формировать коллекции, которые доступны также и для внешних систем.

С активным развитием Semantic Web и его относительно нового направления LOD в сети появляются ресурсы, представляющие огромные объемы информации по разным предметным областям. В число этих ресурсов входят и различные электронные библиотеки. Особая ценность их интеграции в LOD обеспечивается возможностью связать данные из различных источников. Возможность использования тезауруса некоторой предметной области в нашей системе позволяет не просто искать и формировать определенные данные в облаке LOD, но и выявлять новые связи между ними, и дополнять уже имеющиеся данные, опираясь на дополнительные возможности системы.

Итак, мы определяем персональную семантическую электронную библиотеку как информационную систему, в качестве ресурсов которой выступают структурированные коллекции разнородных электронных объектов, за формирование которых отвечают пользователи

системы. Эти объекты поступают в систему из различных источников данных зарегистрированных в системе. Для каждого объекта в системе поддерживается набор соответствующей контекстной информации. Средствами системы поддерживается создание и поддержка тезауруса, который представляет знания о предметной области семантической электронной библиотеки. Основная функциональность системы обеспечивает разнообразные средства навигации и поиска по ресурсам и их источникам, доступным через сеть, а также возможность дальнейшей публикации ресурсов библиотеки в LOD. В процессе подключения и описания новых источников данных, тезаурус пополняется новыми понятиями и связями, расширяя тем самым не только область поиска, но и благодаря связям уточнять и конкретизировать тематику поиска..

Проблеме поиска в LOD посвящены различные исследования и существуют поисковые системы, ориентированные на источники, интегрированные в LOD. В работе [9] описывается система поиска в репозиториях LOD на основе высокоуровневой онтологии, на которую отображается схема подключаемого источника данных. Недостаточный уровень концептуализации понятий не позволяет в достаточной мере сконцентрироваться на определенной предметной области. В системах, описанных в работах [10], [11], требуется знание каждого источника данных для задания поисковых запросов. Поисковые системы, такие как Sig.ma, Falcons и SWSE, обеспечивают поиск на основе ключевых слов. Наш подход конкретизирует предметную область, используя тезаурус в рамках семантической электронной бибилиотеки и позволяя связывать результаты поиска с уже имеющимися ресурсами в репозитории бибилиотеки.

2 Семантический подход к электронным библиотекам

В данной работе не затрагиваются задачи построения онтологии библиографических данных для электронных библиотек. Мы рассматриваем задачу построения онтологии электронной библиотеки как информационной системы. Обычно при построении информационных систем на первом этапе выделяют общие понятия, которые не зависят от конкретной предметной области. Далее вводятся определения, характерные для конкретной предметной области, которые соединяются с общими понятиями бинарными отношениями. Наиболее полная онтология для описания информационных систем (онтология BWW) была представлена в работе [2], которая фокусируется на модели представления, определяет набор понятий, их связей и характеристик, достаточных для описания структуры и поведения информационных систем. Основное преимущество этого подхода - это гибкость и расширяемость систем. Концептуальная модель электронных библиотек, с определениями важнейших представлений об архитектуре, ресурсах и функциональности электронных библиотек была определена в программном документ DELOS Digital Library Reference Model [1].

2.1 Онтология электронной библиотеки

На основе изучения «стандартов» [1], [3] в области электронных библиотек можно сказать, что эта область плохо формализована. Некоторые вопросы хорошо исследованы, но моделей, описывающих компоненты построения электронной библиотеки в целом, не существует, что осложняет построение подобного рода систем. На основе понятий концептуальной модели электронных библиотек предложена онтология системы управления персональной семантической электронной библиотекой. Такая библиотека поддерживает различные профили пользователей (эксперты, администраторы, операторы, простые пользователи) с учетом их прав, предоставляет различные сервисы для работы с контентом (формирование коллекций, рубрикация, активация и деактивация источников данных), поддерживает различные процессы управления подсистемами библиотеки (управление классификаторами, наполнение тезауруса по описанию источников данных и т.д.).

Онтология электронной библиотеки, основные понятия которой представлены ниже, разработана в общем виде без привязки к конкретным методам и способам реализации семантических цифровых библиотек. На рисунке 1 приведен фрагмент UML-диаграммы классов основных сущностей разработанной онтологической модели. В верхней части представлены понятия онтологии BWW, нижняя часть представляет их отображение на понятия электронных библиотек. Этот фрагмент взят из работы [13] и представляет понятия онтологии, связанные с подсистемой управления доступом. Работа [13], основывается на подходе также ориентированном на онтологию BWW, что позволяет нам использовать эту модель для электронной библиотеки для описания роли и прав пользователей при взаимодействии с электронной библиотекой, а также определить правила работы.

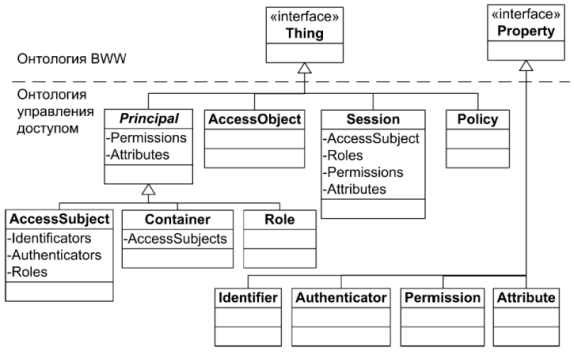

Рисунок 1.

На рисунке 2 приводится фрагмент онтологии подсистемы управления контентом

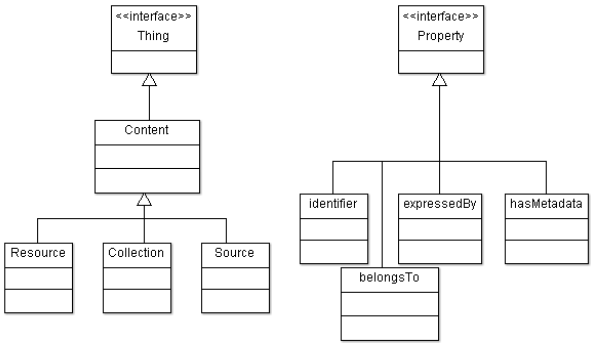

Рисунок 2.

Content (Контент) – это суперкласс объектов электронной библиотеки, задает общие характеристики объектов

Resource (Ресурс) – это информационный объект, множество которых и образует основной контент библиотеки, описание которого дается определенным набором метаданных, представленным соответствующим источником данных

Collection (Коллекция) – подмножество произвольных типов ресурсов

Source (Источник) – представляет собой «параметрическое» описание внешнего по отношению к конкретной библиотеке источника ресурсов (данных) поступающих в систему, где ресурсы могут быть представлены в различных форматах

На рисунке 3 приводится фрагмент онтологии подсистемы управления словарями / классификаторами / тезаурусами

Vocablurary (Словарь) – линейный список терминов

Classifier (Классификатор) – иерархически связанные термины

Taxonomy (Таксономия) - общее представление справочников и словарей

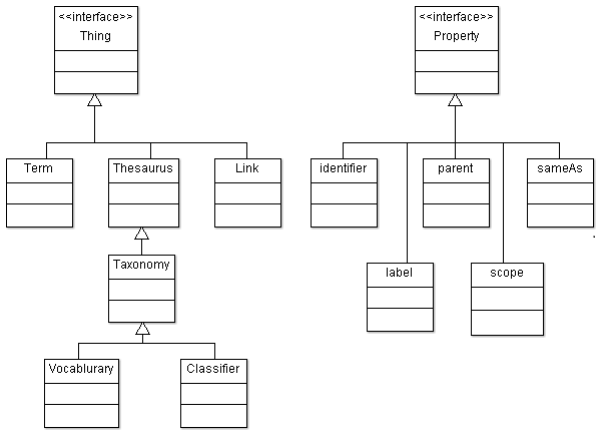

Рисунок 3

Thesaurus (Тезаурус) – является наиболее общей формой таксономии. совокупность словарей/классификаторов, с вертикальными и горизонтальными связями, концепты (элементы) тезауруса могут использоваться как для классификации контента, так и для (описания) источников (ресурсов) данных.

На рисунке 4 приводится фрагмент онтологии подсистемы автоматического мониторинга источников

SavedQuery (Сохраненный запрос) – пользователь определяет запрос к источникам данным, который запрашивает новые объекты за определенный период времени. Запрос определяется с помощью графического интерфейса, то есть от пользователя не требуется специальных знаний, далее системой запрос транслируется в SPARQL и сохраняется в таком виде.

SavedQueryCollection (Коллекция сохраненного запроса) – последняя коллекция новых объектов полученных в результате автоматического мониторинга источников данных

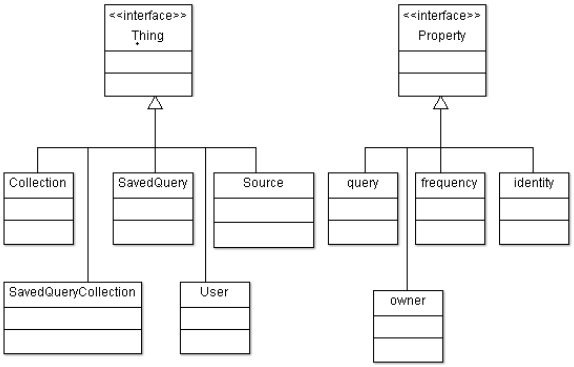

Рисунок 4

3 Общая схема. Основные модули

На рисунке 5 представлена общая схема системы, выделены ее основные модули и дана краткая характеристика каждого.

3.1 Модуль управления доступом

Важной подсистемой любой электронной библиотеки является система управления доступом пользователей к сервисам библиотеки. Пользователь управляет, использует и редактирует контент библиотеки, используя соответствующие доступные сервисы системы. Пользователь должен обладать правами - совокупностью ограничений, накладываемых на него при использовании сервисов системы для работы со своей электронной библиотекой.

3.2 Модуль навигации по ресурсам библиотеки

Подсистема навигации определяет представление данных в различных форматах, обеспечивает навигацию по структуре данных, поддержку тематических подборок, работу с коллекциями объектов, атрибутный поиск, выделение неявных связей между ресурсами по их описаниям.

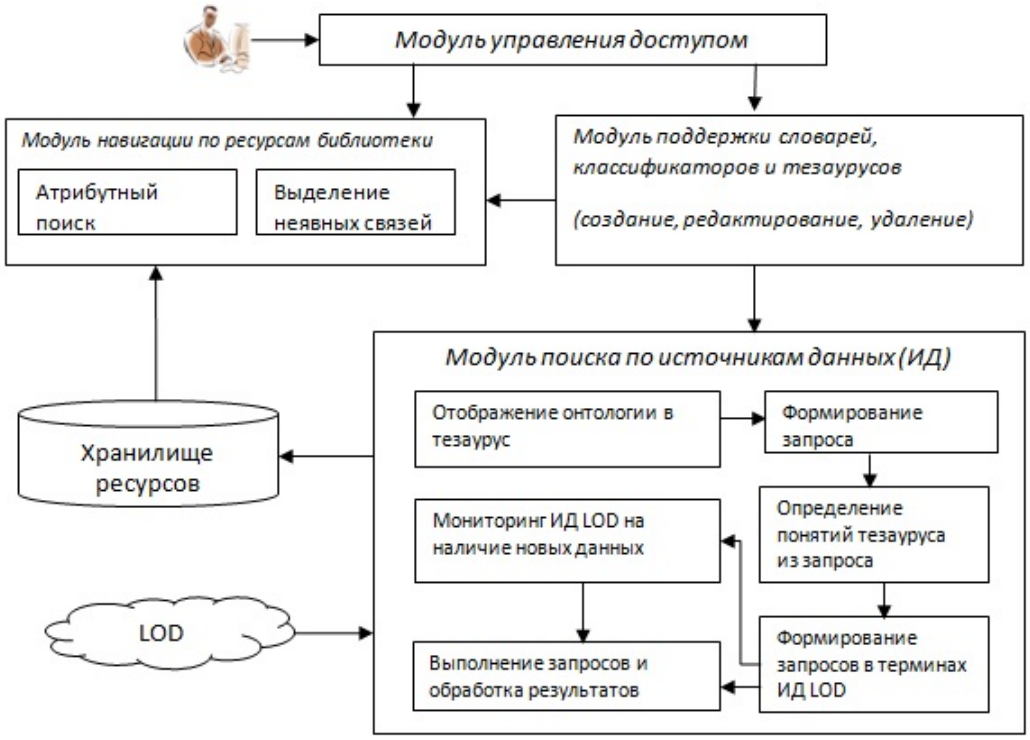

Рисунок 5

3.3 Модуль поиска по источникам данных

Этот модуль позволяет интегрировать данные из внешних систем. Таким образом, контент библиотеки включает не только собственно информационные ресурсы из своего электронного каталога, но и источники данных. Для интеграции данных используется тезаурус, в который отображается описание структуры данных из источников, но при этом тезаурус содержит понятия только той предметной области, в которой заинтересован пользователь. Функция, выполняющая поиск данных во внешних источниках данных, запрашивает описание структур данных из этого тезауруса и отправляет запрос. После этого полученные данные отображаются пользователю. Система может предоставить поиск на естественном языке, при этом пользовательские запросы могут быть проанализированы с помощью функции для трансляции их в запросы для конкретных источников данных. Данный модуль также содержит функцию автоматического мониторинга источников данных, которая информируют пользователей о поступлении новых или изменении существующих информационных объектов в источнике или во внутреннем хранилище, в соответствии с их интересами.

3.4 Модуль поддержки словарей, классификаторов и тезаурусов

Как было отмечено выше, тезаурус обеспечивает представление знания о предметной области семантической электронной библиотеки. Контролируемые словари и классификаторы в системе используются для структуризации данных.

Этот модуль позволяет создавать и наполнять словари, классификаторы, тезаурус, а также позволяет осуществлять просмотр (навигацию) и атрибутивный поиск терминов, чем обеспечивается эффективное выполнение необходимых для этого запросов. Также обеспечиваются функции администрирования тезауруса, при необходимости допускается детализация некоторых связей, а также добавление новых типов связей через интерфейс редактирования плоского словаря этих связей.

Этот модуль поддерживает редактирование набора классификаторов, их структуры и элементов с помощью пользовательских интерфейсов, а также возможность через интерфейсы системы указать перечень классифицируемых типов ресурсов и разорвать связь между некоторым типом ресурсов и классификатором. Каждый классификатор может быть подключен к любому типу ресурсов, и каждый тип ресурсов может классифицироваться несколькими классификаторами.

4 Постороение тезауруса предметной области на основе источников данных

Основная задача системы заключается в предоставлении пользователю унифицированного представления для возможности извлечения интересующей пользователя информации по определенной предметной области. Эта предметная область описывается в терминах тезауруса, поддержка которого обеспечивается соответствующим модулем. При отображении терминов источников данных на этот тезаурус неизбежно возникает основная проблема – неоднородность:

- структурная неоднородность – данные в различных источниках могут быть по-разному представлены и организованы в структуру;
- семантическая неоднородность – данные могут быть представлены в различных системах понятий, схожие понятия могут по-разному интерпретироваться в разных источниках данных.

Для преодоления неоднородности при расширении тезауруса терминами из источников данных необходимо проводить некоторый предварительный анализ понятий источника данных и понятий из тезауруса нашей системы

- на сходство символических имен терминов;
- структурного положения понятия,
- степень сходства множеств атрибутов, достаточных для идентификации объекта и необходимых для его описания, где необходимыми и достаточными атрибутами понятия являются
  - идентифицирующие атрибуты,
  - обязательные описательные атрибуты.

На этом этапе надо уделять внимание анализу связей [4], [6], которые позволяют выявить

- эквивалентные классы,
- ранее не определенные связи между разными источниками данных,
- новые источники данных.

На этапе предварительного анализа актуальны проблемы, классификация которых представлена в работах [7], в которых выделяются следующие группы: лексические, синтаксические, семантические, структурные.

Отображение понятий источников данных в тезаурус производится методом частичного соответствия, где **с**оответствие есть отображение понятий и отношений источника данных на тезаурус [7]. Соответствие может быть определено не полностью и является частичным, если, может существовать несколько понятий в источнике, не имеющих своих эквивалентов в тезаурусе. Для семантического поиска в разных источниках по исследуемой предметной области достаточно такой интеграции на уровне частичного соответствия, которое позволяет избежать изменения в источниках.

При расширении или изменении тезауруса (например, при подключении пользователем нового источника данных для электронной библиотеки) основными операциями являются:

- поиск в тезаурусе понятий эквивалентных понятиям из источника,
- добавление новых понятий из источника в существующий тезаурус,
- привязка к суперпонятиям понятий из источника (если суперпонятия присутствуют в тезаурусе),
- привязка подпонятий к понятиям из источника (если подпонятия присутствуют в тезаурусе).

Основные операции над свойствами:

- добавление новых свойств и связей к понятиям из тезауруса,
- поиск эквивалентных свойств и связей для понятий из источников.

4.1 Источники данных

Источники могут представлять данные трех видов [9]:

- структурированные - предоставляют стандартизированное описание метаданных своих информационных ресурсов, например, в виде онтологий на OWL,
- неструктурированные - не существует каких-либо общепринятых стандартов их представления, но содержат не меньшее количество полезной информации,
- полуструктурированные - обладают некоторой структурой, но не являются жестко структурированными, например, XML.

Недостаточная степень гранулярности структурированных на первый взгляд данных может вызвать затруднения для их обработки, не говоря уже о проблемах в случае с неструктурированными данными. Для преодоления большинства проблем могут использоваться методы text mining. Основная задача text mining - переход от неструктурированного текста к структурированному через последовательность преобразований. Некоторые задачи, которые могут быть решены с их помощью, это:

- выявление связей между ресурсами,
- уточнение тематической направленности,
- выделение ключевых слов.

5 Заключение и дальнейшее направление работы

Сейчас реализовано ядро системы, которое проходит тестовую эксплуатацию. Дальнейшее направление работ планируется в области использования возможностей text mining для анализа сопутствующей текстовой информации и

выявления неявных связей между различными объектами. Эти методы позволят также решать задачи уточнения терминов онтологии предметной области и обработки текстовых документов для более точной их классификации. Реализуется возможность задания запросов на естественном языке к полуструктурированным источникам данных на основе «улучшенной» методами text mining онтологии. Использование методов text mining для уточнения методов построения онтологии предметной области позволит существенно улучшить качество онтологии и позволит соответственно обогащать интегрируемые в системе данные из различных источников, используя более точные понятия и термины, и связи между ними. Также планируется реализовать подсистему позволяющую отслеживать во времени изменение и развитие состояния данных, что позволит оценить «эволюцию» и «распространение» информации по заданной тематике.

Литература

[1] 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. http://www.delos.info/files/pdf/ReferenceModel/DELOS_DLReferenceModel_0.98.pdf

[2] Weber, R. Ontological Foundations of Information Systems , Queensland, Australia, Coopers & Lybrand. 1997.

[3] Functional Requirements for Bibliographic Records, Final Report / IFLA Study Group on the Functional Requirements for Bibliographic Records. – München: K.G. Saur, 1998. (UBCIM Publications, New Series; v. 19) http://archive.ifla.org/VII/s13/frbr/frbr.htm

[4] Lihua Zhao, Ryutaro Ichise. Integrating Heterogeneous Ontology Schema from LOD.

[5] Proceedings of the 26th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI2012), Yamaguchi, June 12-15, 2012.

[6] Ding, L., Shinavier, J., Finin, T., McGuinness, D.L.: An Empirical Study of owl:sameAs Use in Linked Data. In: Web Science 2010.

[7] Erhard Rahm, Hong Hai Do. Data Cleaning: Problems and Current Approaches. 2000

[8] Isabel F. Cruz and Huiyong Xiao, The Role of Ontologies in Data Integration, Journal Of Engineering Intelligent Systems, 2005, volume 13, pages 245—252.

[9] М. Р. Когаловский. Метаданные, их свойства, функции, классификация и средства представления RCDL, 2012. 3-14

[10] Jain, P., Verma, K., Yeh, P.Z., Hitzler, P., Sheth, A.P.: LOQUS: Linked Open Data SPARQL Querying System. Technical report, Tech. rep., Kno. e. sis Center, Wright State University, Dayton, Ohio, 2010. Available from http://www.pascal-hitzler.de/resources/publications/loqus-tr-2010. pdf (2010)

[11] Hartig, O.; Bizer, C.; and Freytag, J.-C. 2009. Executing SPARQL Queries over the Web of Linked Data. In ISWC 2009 , volume 5823 of LNCS , 293–309

[12] Quilitz, B., and Leser, U. 2008. Querying Distributed RDF Data Sources with SPARQL. In ESWC 2008 , volume 5021 of LNCS, 524–538.

[13] Созыкин А.В. Семантическая интеграция управления доступом к сервисам. Интернет ресурс. – Режим доступа : www/URL: http://asozykin.ru/sites/default/files/sozykin.pdf

An Approach to Creating a Personal Semantic Digital Library

O.M. Atayeva, V.A. Serebryakov

The aim of this work is to develop an information system for creation of semantic digital library which content is individual for each user and is populated from various data sources located on the Web and integrated into LOD cloud. We specify its main modules and provide brief characteristic of each of the modules. The system gives to the user a unified presentation in order to enable retrieval of the information on a certain subject area the user is interested in. This subject area is described in terms of the thesaurus supported by a respective module. The data source terms mapping onto this thesaurus is discussed.