OSTIS-2016
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

СЕМАНТИЧЕСКИЕ АСПЕКТЫ ПРЕДСТАВЛЕНИЯ МЕТАДАННЫХ В СИСТЕМЕ «КОРПУС-МЕНЕДЖЕР»

Невзорова О.А., Мухамедшин Д.Р., Курманбакиев М.И.

НИИ «Прикладная семиотика» Академии Наук Республики Татарстан, г. Казань, Россия

onevzoro@gmail.com

damirmuh@gmail.com

write@marat.link

В статье рассказывается об опыте разработки модели представления метаданных в системе управления лингвистическими данными (корпус-менеджер), предназначенной для работы с электронным корпусом текстов татарского языка. В частности, в работе затронуты семантические аспекты представления метаданных.
Ключевые слова: метаданные, корпус-менеджер, семантическая сеть, RDF.

Введение

Одним из важных аспектов эффективной работы с электронным корпусом текстов является корректное описание так называемой «внешней» разметки метаданных текстов, которая включает библиографические данные, а также тематические, социологические и типологические характеристики документов. Наиболее полное описание метаданных поможет расширить возможности поискового механизма корпуса, что в свою очередь позволит расширить область различных исследований по изучению языка с использованием корпуса.

Для представления метаданных в текстовых документах разработан ряд международных стандартов и рекомендаций. Наиболее ранние рекомендации связаны с представлением метаданных в виде семейства форматов библиографических записей MARC, в частности его «русифицированная» версия RUSMARC. Данный формат позволяет довольно широко описывать текстовые данные (книги, журналы и пр.). Но использование форматов MARC при каталогизации изданий требует специальной подготовки, и, к сожалению, методические рекомендации, при использовании данного стандарта, довольно часто полностью не соблюдаются.

Специальные рекомендации EAGLES Preliminary recommendations on text typology (EAG-TCWG - TTYP/) разработаны для представления метаданных именно для использования в электронных корпусах Консультативной группой экспертов по техническим стандартам языка (Expert Advisory Group on Language Engineering Standards, EAGLES). Данные рекомендации наиболее полно отражают объективный набор метаданных, необходимый для представления данных электронного корпуса и его потенциальных приложений.

В рекомендациях выделено два класса факторов метаданных, влияющих на язык текстов: внешний и внутренний [Sinclair, 1996].

К внешним факторам относят:

- E.1. Origin – факторы, касающиеся происхождения текста, которые считают оказавшими влияние на структуру или содержание текста, такие как автор текста, его переводчик, правообладатель, язык оригинала текста, время написания и прочее.
- E.2. State – факторы, касающихся внешнего вида текста, его расположение и отношение относительно нетекстовых материалов, такие как тип носителя текста, отношения с нелингвистической коммуникативной средой (изображения, диаграммы и прочие нетекстовые элементы), аспекты в области дизайна, оказавшие влияние на язык и прочее.
- E.3. Aims – факторы, касающиеся причин создания текста и его влияния на аудиторию, такие как размер аудитории, профессиональная область аудитории, связи автора с аудиторией, тип текста и прочее.

К внутренним факторам относят:

- I.1. Topic – предметная область текста
- I.2. Style — стилистические особенности текста.

Удобным вариантом представления модели метаданных является концепция Дублинского Ядра (Dublin Core) [Hillman, 2005], предназначенная для унификации метаданных широкого диапазона ресурсов. Простой набор Дублинского ядра (Dublin Core Metadata Element Set; DCMES) включает в себя 15 элементов: contributor, coverage, creator, date, description, format, identifier, language, publisher, relation, rights, source, subject, title, type [DCMES, 2012]. Данные 15 элементов были включены в стандарты IETF RFC 5013 (RFC5013) /(2007), ISO Standard 15836:2009 (ISO15836) / (2009), ANSI/NISO Standard Z39.85-2012 (NISOZ3985) / (2013). Квалифицированный набор данных, помимо перечисленных выше 15 элементов, может включать в себя дополнительно 54 элемента [DCMI Metadata Terms, 2012].

DCMI представлен в форматах DC-TEXT, HTML, XML, RDF. В настоящий момент Dublin Core активно развивается, разрабатываются новые элементы описания метаданных, расширяется сфера задач, в которых применим данный словарь.

В настоящей статье предложена достаточно полная модель описания метаданных в электронном корпусе, построенная с учетом рекомендаций EAGLES и DCMI.

1. Прототип корпус-менеджера

В настоящее время разработаны различные по функционалу системы «корпус-менеджеры», предназначенные для решения определенного круга задач. Можно указать ряд наиболее актуальных разработок для целей настоящей статьи. Восточно-армянский национальный корпус (http://eanc.net), на платформе которого располагается электронный корпус татарского языка «Туган Тел» (http://web-corpora.net/TatarCorpus), имеет 5 различных видов поиска: прямой поиск по слову, прямой поиск по лемме, обратный поиск, точный и неточный поиск. Функционал, реализованный в платформе Восточно-армянского национального корпуса, является базовым для корпус-менеджера, представленного в настоящей статье.

Национальный корпус русского языка (http://ruscorpora.ru) обладает функционалом, схожим с платформой Восточно-армянского национального корпуса. Отличительной особенностью платформы Национального корпуса русского языка является поддержка расширенного синтаксиса поисковых запросов при прямом поиске, а именно поддержка минус-слов, поиска по части слова и логических операторов.

На основе открытых решений авторами разработан новый корпус-менеджер [Невзорова и др., 2015а; Невзорова и др., 2015б; Nevzorova et al., 2015]. Функционал системы в основных функциях соответствует функционалу платформ Национального корпуса русского языка и Восточно-армянского национального корпуса, но реализация конкретных задач позволяет говорить о существенном приросте эффективности данной платформы как со стороны обширности функционала, так и со стороны скорости взаимодействия пользователя с системой. Основными преимуществами корпус-менеджера, разработанного авторами, также являются готовая поддержка татарского языка и возможность быстрой интеграции с электронными корпусами других языков, в первую очередь, тюркских, поддержка произвольных морфологических формул, выявление логических ошибок, открытость используемых технологий.

Помимо разработки расширенного поискового функционала перед авторами стояли задачи оптимизации времени исполнения поисковых запросов (менее 1 секунды на запрос), поддержка произвольных морфологических формул с использованием операторов И, ИЛИ, НЕ и выставления приоритетов выполнения при помощи скобок, а также выявление логических ошибок в формулах. Примером логической ошибки является противоречивая формула «!(N|V),INF_1», которая означает «НЕ имя существительное (N) И НЕ глагол (V) И инфинитив, оканчивающийся на аффикс -*ырга* (INF_1)». Противоречивость данной формулы заключается в том, что все элементы, относящиеся к классу «INF_1» также относятся и к классу «V», но в первой части все элементы класса «V» исключаются, соответственно, результатов по данному поисковому запросу существовать не может.

2. Метаданные в корпус-менеджере

Исходными данными для системы корпус-менеджер являются текстовые документы с морфологической разметкой, которая автоматически производится с использованием модуля двухуровневого морфологического анализа татарского языка, реализованного в программном инструментарии PC-KIMMO.

Объем татарского корпуса на декабрь 2015 года составляет более 82 млн словоформ. Корпус содержит тексты различных жанров (художественная литература, тексты СМИ, тексты официальных документов, учебная литература, научные публикации и др.) [Suleymanov, 2013].

В соответствии с рекомендациями EAGLES в наборе метаданных для описания текстового документа в электронном корпусе выделяются три блока (внешние факторы, внутренние факторы и технические метаданные).

Внешние факторы:

- Тип текста (оригинал или перевод);
- Название;
- Автор;

- Переводчик;
- Издание;
- Издательство;
- Язык;
- Дата создания;
- Объем в словах;
- Количество слов в русском файле;
- Источник оригинала;
- Источник перевода;
- Ключевые слова;
- Информация об авторском праве;
- Краткое описание;
- Примечание.

Внутренние факторы:

- Стиль;
- Категория/тематика;
- Место.

Технические метаданные:

- Номер (ID);
- Название исходного файла;
- Русский файл;
- Флаг проверки данных модератором.

Для представления документа в корпусе необходимо, чтобы этот документ имел обязательный (минимальный) набор метаданных, относящийся к внешним факторам, а именно:

- Название;
- Объем в словах;
- Номер (ID);
- Имя исходного файла.

Семантика документа представляется посредством внутренних факторов, таких как узкая тематика текста, хронотоп и др. Целью будущих исследований является разработка методов автоматического определения значений внутренних факторов метаданных, что позволит использовать корпусные документы в широком классе приложений, связанных с обработкой текстов.

3. Модель представления метаданных в корпус-менеджере

Так как метаданные документов могут состоять из неограниченного числа свойств и в будущем дополняться новыми данными, модель представления этих метаданных должна обеспечивать полноту и масштабируемость. Универсальным решением задачи, стоящей перед авторами, является использование семантической сети, представленной в модели представления данных RDF. Таким образом, информационной моделью метаданных корпуса является семантическая сеть. В качестве основополагающей модели было принято решение использовать рекомендации DCMI [DCMI Metadata Terms, 2012], в которых описаны все метаданные, присутствующие у существующих документов электронного корпуса.

Использование модели RDF обусловлено в первую очередь тем, что модель данных, построенную на основе модели RDF можно очень просто масштабировать в ширину, добавляя новые объекты и связи, но не затрагивая основную массу данных. Именно простота модели RDF помогает строить понятные запросы к БД даже при высокой сложности поискового запроса и использовать при этом конечное количество таблиц, о чем говорится ниже в этой статье.

Часто бывает, что у документа отсутствует часть метаданных и её невозможно получить из текста документа. В этих случаях одним из возможных решений может быть использование внешних источников данных, таких как Open Library [openlibrary.org], Computer Science Bibliography, библиотека конгресса США [loc.gov]. Использование модели RDF может помочь в связывании этих внешних данных с элементами метаописания документов.

Такой подход к тому же позволит описать семантику выходных данных внутри HTML-документа, не теряя полноты данных, при помощи рекомендаций RDFa.

Заполнение метаданных является одной из самых ресурсоемких задач. В связи с этим авторами разрабатываются методы автоматизации этого процесса путем получения метаданных из документа и использования внешних источников данных. Решение данной задачи и является основной целью разработки структуры метаданных, которая позволила бы автоматизировать заполнение метаданных, исключить ошибки, возникающие из-за человеческого фактора.

Список свойств метаописания, представленный выше, позволяет определить основные объекты и связи между ними для представления метаданных. Они показаны на Рисунке 1 и в Таблице 1.

Таблица 1 – Основные объекты модели метаданных в корпус-менеджере

Свойства	Связь	Тип данных/ Класс
Номер (ID)	hasID	Число
Исходный файл, Объем в словах	hasSourceFile	SourceFile
Тип текста	hasType	Строка
Язык	hasLanguage	Строка
Название	hasName	Строка
Автор	hasAuthor	Author
Переводчик, Русский файл, Количество	hasTranslation	Translation

слов в русском файле, Источник перевода		
Стиль	hasStyle	Строка
Дата создания	hasCreationDate	Дата
Источник оригинала	hasSource	Строка
Издание, Издательство	hasEdition	Edition
Категория/ тематика	hasCategory	Category
Место	hasPlace	Place
Ключевые слова	hasKeyword	Строка
Информация об авторском праве	hasCopyright	Строка
Краткое описание	hasDescription	Строка
Примечание	hasNote	Строка
Флаг проверки данных модератором	isChecked	Логический

Некоторые метаданные корпуса Татарского языка объединяются в классы и имеют свойственные только им связи с объектами:

- SouceFile (Исходный файл):
  - hasFilename («имеет имя файла»);
  - hasContentSize («имеет размер содержимого»);
- Person (Человек):
  - hasName («имеет имя»);
  - hasSurname («имеет фамилию»);
  - hasMiddleName («имеет отчество»);
  - hasPseudonym («имеет псевдоним»);
  - hasWritingVariant («имеет вариант написания»);
  - hasBirthDate («имеет дату рождения»);
  - hasDeathDate («имеет дату смерти»);
  - hasBirthplace («имеет место рождения»);
- Organization (Организация):
  - hasName («имеет название»);
  - hasLegalName («имеет юридическое название»);
  - hasAddress («имеет адрес»);
  - hasWebsite («имеет веб-сайт»);
  - hasEmailAddress («имеет адрес электронной почты»);
  - hasFoundingDate («имеет дату основания»);
- Author (Автор):
  - isA («является», субъектом для связи является объект класса Person или Organization);
- Translation (Перевод):
  - hasTranslator («имеет переводчика», субъектом для связи является объект класса Person или Organization);
  - hasSourceDocument («имеет исходный документ», применим, если исходный документ является элементом множества Документы);
  - hasSourceFile («имеет исходный файл»);
- Edition (Издание):
  - hasPublishingHouse («имеет издательство», субъектом для связи является объект класса Organization);
  - hasPublishingDate («имеет дату издательства»);
- Category (Категория):
  - hasName («имеет название»);
  - hasParentCategory («имеет родительскую категорию», субъектом для связи является объект класса Category);
- Place (Место):
  - hasType («имеет тип», например, физическое или вымышленное);
  - hasName («имеет название»);
  - hasAddress («имеет адрес»);
  - hasCoordinates («имеет координаты»).

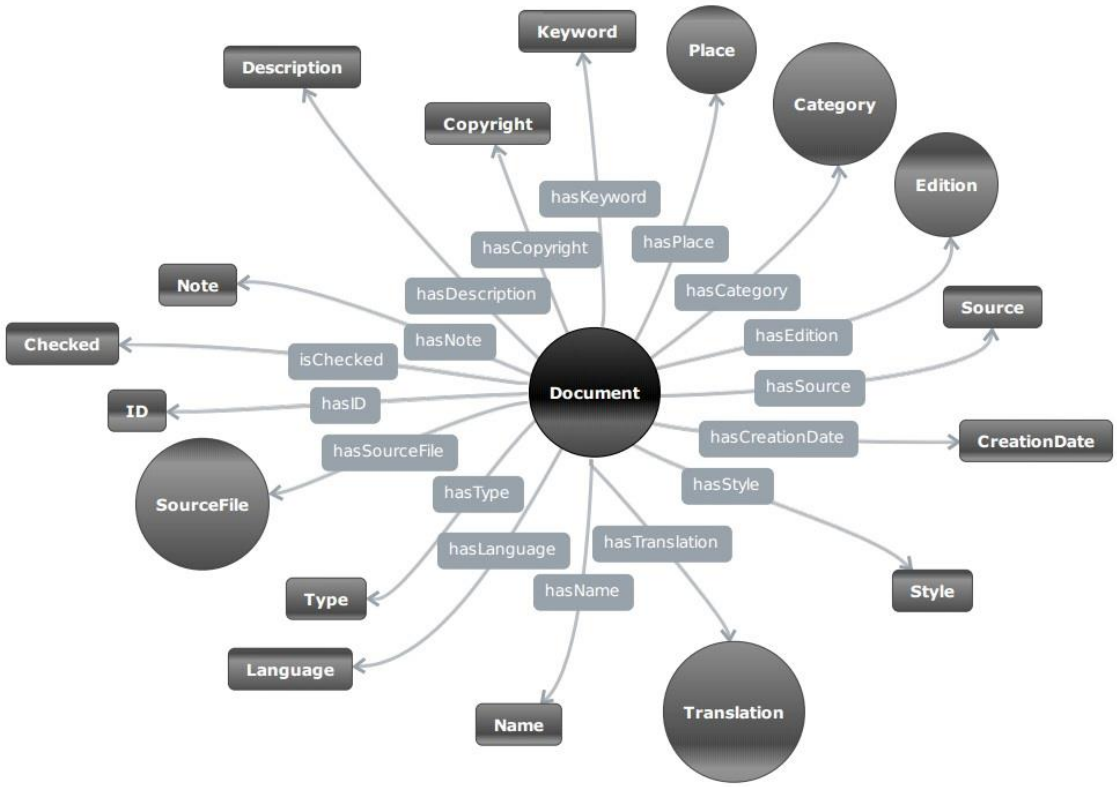

Рисунок *1* – Структура метаданных в корпус-менеджере

4. Технические аспекты представления метаданных

Для функционирования корпус-менеджера используется следующее открытое программное обеспечение: веб-сервер Apache (или HHVM), интерпретатор PHP, СУБД MariaDB, in-memory хранилище Redis (кэширующий сервис), сервер очереди MemcacheQ. Всё программное обеспечение распространяется с открытым исходным кодом и может быть свободно использовано в некоммерческих целях.

Метаданные документов и сами документы хранятся в БД MariaDB, эта БД является реляционной, поэтому для хранения в ней данных, представляемых в модели RDF достаточно иметь одну таблицу *metadata* с полями *subject*, *predicate*, *object*. Эти три поля соответствуют стандарту RDF.

Для того, чтобы оптимизировать скорость выполнения запросов на выборку, в полях таблицы *metadata* хранятся лишь идентификаторы соответствующих субъектов, предикат и объектов. Сами субъекты, предикаты и объекты хранятся в отдельных таблицах и связаны с таблицей *metadata*.

Так, таблица *subjects* имеет поля *id*, *subject* и *type*, таблица *predicates* – *id*, *predicate*, таблица *objects* – *id*, *object*, *type*. К тому же в БД существует таблица *types*, которая содержит в себе типы субъектов и объектов для правильной работы с данными. Таблица *types* имеет поля *id*, *type*.

Заключение

Предложенная в данной работе модель представления метаданных реализована в модуле управления контентом системы управления лингвистическими данными, работающей с электронным корпусом татарского языка.

Данная модель позволяет масштабировать метаданные без ущерба для существующих данных. При этом модель предоставляет возможность максимально полно описывать документы электронного корпуса. Использование семантической сети и модели представления данных RDF в качестве основы для модели представления метаданных в корпус-менеджере позволяет не только оптимизировать работу поискового модуля и находить документы по любому свойству, но и описывать семантику выходных метаданных непосредственно в HTML-документе. Эти и другие преимущества модели представления метаданных, описанной авторами в данной статье, позволяют говорить о том, что выбранное направление является оптимальным при решении поставленной задачи.

Использование семантических сетей для представления данных лингвистических корпусов позволяет покрыть широкий спектр задач. В статье описаны лишь некоторые из них. В перспективе предполагается использование этой же схемы для реализации морфологического анализатора и решения задачи снятия омонимии.

Общий подход к решению задач в поисковой системе по лингвистическому корпусу может

позволить использовать эту систему не только с электронным корпусом текстов на татарском языке, но и с корпусами других языков, не прибегая к значительным изменениям в системе.

На данный момент активно ведется изучение возможностей заполнения метаданных внутреннего фактора с помощью полуавтоматического или автоматического семантического аннотирования. Предполагается реализация данной возможности с помощью методов идентификации именованных сущностей и отнесения их к определенному семантическому классу.

Библиографический список

[Невзорова, 2015] Невзорова О.А., Мухамедшин Д.Р., Билалов Р.Р. Семантические аспекты представления и обработки поисковых запросов в системе корпус-менеджер // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2015): материалы II Междунар. научн.-техн. конф. (Минск, 19-21 февраля 2015 г.) / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск: БГУИР, 2015. – С. 439-444.

[Невзорова, 2015] Невзорова О.А., Мухамедшин Д.Р., Билалов Р.Р. Корпус-менеджер для тюркских языков: основная функциональность // Труды международной конференции «Корпусная лингвистика - 2015». – СПб.: С.-Петербургский гос. Университет, филологический факультет, 2015. – С. 344-350.

[Suleymanov, 2013] Dzhavdet Suleymanov, Olga Nevzorova, Ayrat Gatiatullin, Rinat Gilmullin, Bulat Khakimov National corpus of the Tatar language “Tugan Tel”: Grammatical Annotation and Implementation // Procedia - Social and Behavioral Sciences (2013), pp. 68-74.

[Nevzorova, 2015] Nevzorova O., Mukhamedshin D., Bilalov R. Search Engine for the 'Tugan Tel' Tatar National Corpus: Main Decisions // Proceedings of the International Conference “Turkic Languages Processing” TurkLang-2015 — Kazan, 2015. - Pp. 236-244.

[Sinclair, 1996] J. McH. Sinclair, J. Ball. EAGLES Preliminary Recommendations on Text Typology EAG---TCWG---TTYP/P: [Электронный ресурс]. 1996. URL: http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html (Дата обращения: 18.12.2015)

[Hillman, 2005] Diane Hillmann. Using Dublin Core: [Электронный ресурс]. 2005. URL: http://dublincore.org/documents/usageguide/ (Дата обращения: 25.11.2015)

[DCMES, 2012] Dublin Core Metadata Element Set, Version 1.1 [Электронный ресурс]. 2012. URL: http://dublincore.org/documents/dces/ (Дата обращения: 25.11.2015)

[DCMI Metadata Terms, 2012] DCMI Metadata Terms [Электронный ресурс]. 2012. URL: http://dublincore.org/documents/dcmi-terms/ (Дата обращения: 25.11.2015)

SEMANTIC ASPECTS OF METADATA REPRESENTATION IN CORPUS MANAGER SYSTEM

Nevzorova O.A., Mukhamedshin D.R., Kurmanbakiev M.I.

Research Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan, Russia

onevzoro@gmail.com

damirmuh@gmail.com

write@marat.link

The article tells about the experience of the development model of metadata representation in linguistic data management system (corpus manager), which is designed to work with corpus of texts of the Tatar language.

Introduction

At present time there are different recommendations (or standards) for annotating text document. In this paper we propose the structure of the document metadata based on the well-known standards such as EAGLES and DCMI.

Main Part

The volume of the Tatar corpus is more than 82 million word forms in June 2015. The corpus contains the texts of different genres (belles-lettres, mass media texts, texts of official documents, educational literature, scientific publications, etc.). Each corpus document has a meta-description: ID; the name of the source file; type of text (original or translation); language; name; author; translator; style; date of creation; the amount of words the source of original; the source of translation; edition; Publishing house; category/theme; place; keywords; information about copyright; short description and others.

Since the metadata can consist of an unlimited number of properties and will be complemented with new data in the future, the model of metadata representation must ensure completeness and scalability. A universal solution of the problem is to use a semantic network, represented in the RDF model of data.

List of properties of meta-description presented above allows you to define the main objects and relationships between them to represent metadata (see Table 1 and the list below).

Metadata of documents and the documents themselves are stored in the MariaDB database. This is relational database, so for storing therein the data submitted in the RDF model is sufficient to have one table ‘metadata’ with fields ‘subject’, ‘predicate’ and ‘object’. These three fields conform to the RDF standard.

Conclusion

The model of metadata representation proposed in this paper is implemented in content management module of linguistic data management system, which works with electronic corpus of the Tatar language.

The use of semantic networks for data representation in linguistic corpora allows covering a wide spectrum of problems. This article describes a few of them. In the future, it is planned to use the same scheme for the implementation of the morphological analyzer and solving the problem of disambiguation.

A general approach to solving problems in the search engine for linguistic corpora may allow the use of this system not only for Tatar corpus, but also for another corpora (another languages) without significant changes in the system.