8 ЭКОНОМИКА И ФИНАНСЫ. МЕНЕДЖМЕНТ

УДК 004.822

РАЗРАБОТКА ОНТОЛОГИИ ОТКРЫТЫХ ГОСУДАРСТВЕННЫХ ДАННЫХ НА ПРИМЕРЕ РАСХОДОВ БЮДЖЕТА САНКТ-ПЕТЕРБУРГА

О.В. Пархимович, В.В. Власов, Д.И. Муромцев

Представлено описание методологии и процесса разработки онтологии открытых государственных данных на примере ведомственной структуры расходов бюджета Санкт-Петербурга. Выполнено обоснование необходимости в создании онтологии как с точки зрения формирования единого терминологического словаря для пользователей и программных агентов, так и для интеграции информационных ресурсов и систем. Приведена общая структура онтологии, описан процесс ее реализации. Кроме того, предложенная схема разработки онтологии позволяет осуществлять ее дальнейшую модификацию и дополнение при изменении исходных данных или нормативных документов.
Ключевые слова: открытые государственные данные, онтологии, семантические сети.

Введение

В последнее время проблематика создания и публикации открытых государственных данных становится более актуальной: например, со слов экс-министра экономического развития Российской Федерации (РФ) Эльвиры Набиуллиной, «Минэкономразвития России считает необходимым развивать тематику открытых государственных данных и планирует проводить с 2012 года ряд работ в рамках государственной программы «Информационное общество (2011–2020 годы)» [1].

Под открытыми государственными данными в мировой практике понимают публичную государственную информацию, предоставляемую в цифровом виде посредством сети Интернет в форме, допускающей последующий анализ и ее повторное использование. В результате становится возможным создание сервисов и проектов, решающих отдельные проблемы граждан, появление новых возможностей для коммерческой деятельности, подкрепляющей экономику государства, повышение эффективности работы государственных органов и увеличение прозрачности управления государством.

В работе отражены результаты разработки онтологии открытых государственных данных ведомственной структуры расходов бюджета (ВСРБ) Санкт-Петербурга, которая является одним из Приложений Закона Санкт-Петербурга о бюджете [2], содержащим распределение бюджетных ассигнований, предусмотренных Законом о бюджете на соответствующий финансовый год главным распорядителем бюджетных средств по разделам, подразделам, целевым статьям и видам расходов бюджетной классификации РФ. Выбор именно этих массивов данных обусловлен их доступностью (данные публикуются на сайте администрации Санкт-Петербурга), достоверностью, актуальностью и наличием четкой структуры данных, что значительно упрощает проведение экспериментальной и аналитической части работы.

В настоящее время Приложения к Закону о бюджете Санкт-Петербурга, содержащие данные о бюджете, публикуются в формате PDF [3], который не содержит структуры информации и не может быть использован для анализа и машинной обработки (например, для автоматизированной визуализации данных). Важность публикации в структурированном формате информации о бюджете городов подтверждается открытием в октябре 2011 г. портала «Открытый бюджет» города Москвы [4], на котором можно не только получить всю интересующую информацию в простой и доступной форме (как текстовой, так и визуальной), но и скачать массивы данных в структурированном формате, а также предложить свои идеи по изменению структуры бюджета.

В ВСРБ Санкт-Петербурга содержится важная информация о распределении доходов города. Значительная часть доходов формируется из налоговых отчислений граждан, поэтому они должны иметь возможность изучения и анализа рассматриваемого документа в удобных для них форматах. Данная работа позволяет решить указанную проблему, предоставляя указанные данные в формате онтологии, пригодном для машинной обработки, что делает анализ данных менее трудоемким.

Постановка задачи

Онтология расходов бюджета Санкт-Петербурга предназначена для предоставления информации в структурированных форматах, которые могут быть проанализированы пользователями или использованы разработчиками при создании приложений.

Структура разработки онтологии представлена на рис. 1 и заключается в следующем. Массив данных в формате PDF скачивается с официального источника. После этого исследуется его структура, на основании которой проектируется иерархия классов онтологии. Полученная иерархия классов создается в системе Protégé и заполняется экземплярами классов. Полученный проект системы Protégé экспортируется в формат RDFs в виде двух RDFs-файлов: файла иерархии классов и файла экземпляров.

Рис. 1. Схема разработки онтологии

Исследование структуры данных ВСРБ Санкт-Петербурга

Термин «ведомственная структура расходов бюджета» в Бюджетном кодексе РФ определен как распределение бюджетных ассигнований, предусмотренных законом (решением) о бюджете на соответствующий финансовый год главным распорядителем бюджетных средств, по разделам, подразделам, целевым статьям и видам расходов бюджетной классификации РФ [5].

ВСРБ Санкт-Петербурга представляет собой таблицу, состоящую из сумм и статей расходов и классификаторов, обеспечивающих международную сопоставимость данных. Коды содержат основную информацию, необходимую для анализа бюджета. Чтобы перевести данные в структурированный формат, необходимо сгенерировать справочники классификаторов, благодаря которым станет возможным автоматизированный анализ информации. Согласно Указаниям «О порядке применения бюджетной классификации Российской Федерации», утвержденным Приказом Министерства финансов РФ от 28.12.2010 № 190н, «классификация расходов бюджетов представляет собой группировку расходов бюджетов всех уровней и отражает направление бюджетных средств на выполнение единицами сектора государственного управления и местного самоуправления основных функций, решение социально-экономических задач» [5]. В структуре расходов используются следующие классификаторы.

- Коды главных администраторов доходов бюджета. Главными администраторами доходов являются органы государственной власти Санкт-Петербурга, которые имеют право распределять бюджетные средства между подведомственными распорядителями и получателями бюджетных средств.
- Код раздела и подраздела. В классификации расходов бюджета содержатся 14 разделов, отражающих направление финансовых ресурсов на выполнение основных функций государства. Каждый раздел детализирован подразделами, конкретизирующими направление бюджетных средств.
- Код целевой статьи. Целевые статьи обеспечивают привязку бюджетных ассигнований к конкретным направлениям деятельности субъектов бюджетного планирования и участников бюджетного процесса в пределах подразделов классификации расходов бюджетов. Код целевой статьи состоит из 7 знаков: первые три знака отражают статью расходов, следующие два – код программы (принадлежность расходов к соответствующему закону, иному нормативно-правовому акту, устанавливающему выплату), и последние – код подпрограммы (виды выплат в рамках закодированного на уровне программы закона, иного нормативно-правового акта). Перечни целевых статей, используемых в бюджетах субъектов РФ, формируются соответствующим финансовым органом субъекта РФ.
- Код вида расходов. Виды расходов детализируют направление финансирования расходов бюджетов как по целевым статьям, так и по целевым программам расходов. Они утверждаются федеральным законом на соответствующий год.
- Код операций сектора государственного управления (ОСГУ). Классификация ОСГУ является группировкой операций в зависимости от их экономического содержания. В рамках данной классификации ОСГУ разделены на текущие (доходы и расходы), инвестиционные (операции с нефинансовыми активами) и финансовые (операции с финансовыми активами и обязательствами). Группы детализируются статьями и подстатьями.

Проектирование онтологии

Онтологии необходимы для определения словаря терминов, совместно используемых специалистами. Они состоят из машинно-интерпретируемых формулировок основных понятий и отношений между ними, определяя общий словарь для ученых, которым необходимо использовать информацию о предметной области. Онтологии обычно используются другими приложениями, например, онтология ВСРБ Санкт-Петербурга может применяться для визуализации содержащихся в ней данных по выбранным пользователем параметрам. Одной из основных задач онтологий является совместное использование людьми или программными агентами общего понимания структуры информации.

В рассматриваемом случае онтология – формальное явное описание понятий в рассматриваемой предметной области (классов), слотов (свойств каждого понятия) и наложенных на них ограничений [6]. Класс является центральным элементом онтологии и описывает понятия предметной области. Например, класс главных администраторов доходов представляет всех администраторов, а конкретные администраторы, например, «Администрация Фрунзенского района» являются экземплярами этого класса. У класса могут быть подклассы, представляющие более конкретные понятия. Например, класс администраторов доходов можно разделить на подкласс администраций районов Санкт-Петербурга и подкласс комитетов

Санкт-Петербурга. Слоты описывают свойства классов и экземпляров: у класса администраторов доходов могут быть слоты «код» и «руководитель».

Перед проектированием онтологии необходимо определить, какую область она будет охватывать и ответы на какие типы вопросов в ней должны содержаться. В данном случае областью онтологии является представление структуры ВСРБ Санкт-Петербурга и сумм расходов. Используя приложения, основанные на данной онтологии, пользователь сможет получать информацию по интересующим его статьям расходов, сравнивать расходы различных распорядителей бюджетных средств или суммы по операциям сектора государственного управления. Разработку онтологии можно разбить на три этапа: определение классов и их иерархии, определение слотов и описание допустимых значений, создание экземпляров классов и заполнение значений слотов экземпляров.

В иерархии классов ВСРБ Санкт-Петербурга необходимо выделить классы используемых в бюджете классификаторов (распорядители бюджетных средств, операции сектора государственного управления, разделы и др.) и класс расходов, содержащий значения сумм расходов с их классификаторами.

После определения классов создается внутренняя структура понятий, которая описывается с помощью свойств классов. Свойства могут быть внутренними (например, количество сотрудников распорядителей бюджетных средств), внешними (например, государственный орган, контролирующий расходы распорядителей бюджетных средств), частью целого (например, функциональные подразделения распорядителей бюджетных средств), отношениями с другими индивидуальными концептами. Слоты могут иметь различные ограничения, описывающие тип значения, разрешенные значения, число значений (мощность) и другие свойства значений. Например, значение слота «название» в классе «распорядители бюджетных средств» – слот с типом значения «строка». Мощность слота определяет количество значений, которые может иметь слот. Например, слот «сумма» класса «Расходы» может иметь только одно значение, т.е. единичную мощность.

Последним шагом в разработки онтологии является создание экземпляров классов. Для определения отдельного экземпляра необходимо выбрать класс, создать отдельный экземпляр этого класса и ввести значения слотов. Например, можно создать отдельный экземпляр класса «Операции сектора государственного управления», значением слота «Наименование» которого будет строка «Оплата труда и начисления на выплаты по оплате труда», а значением слота «Код ОСГУ» будет число «210». Благодаря созданию классов онтологии создается структура расходов бюджета Санкт-Петербурга, а в результате создания отдельных экземпляров эта структура наполняется конкретными значениями.

На основании результатов исследования структуры данных ВСРБ Санкт-Петербурга была разработана онтология, структура которой представлена на рис. 2. В данной онтологии содержатся 6 классов с классификаторами, применяемыми в бюджете Санкт-Петербурга (классы «Главный распорядитель бюджетных средств», «Вид расходов», «Операции сектора государственного управления», «Подраздел», «Раздел», «Целевая статья») и класс «Расходы», содержащий все строки с суммами расходов ВСРБ Санкт-Петербурга.

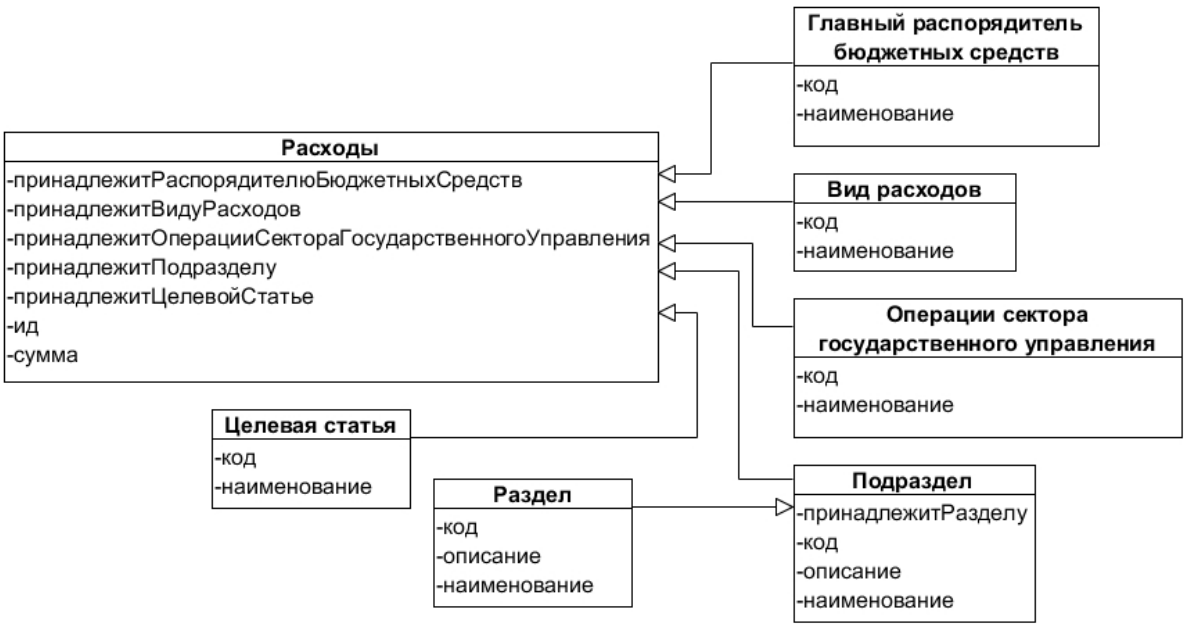

Рис. 2. Структура онтологии ВСРБ Санкт-Петербурга

Реализация онтологии в системе Protégé

На основе структуры классов для ВСРБ Санкт-Петербурга разработана онтология с помощью программного обеспечения Protégé [7]. Выбор системы Protégé для реализации онтологии объясняется тем, что данная программа является свободным, открытым редактором онтологий и фреймворком для построения баз знаний. Онтологии, разработанные с помощью данной системы, могут быть экспортированы в форматы RDF (RDF Schema), OWL и XML-схема. В работе использован редактор Protégé-OWL, позволяющий пользователям строить онтологии для семантической паутины, в частности на OWL [8]. С помощью такой онтологии формальная семантика OWL определяет, как получать логические следствия (факторы, которые не присутствуют непосредственно в онтологии, но могут быть выведены из существующих посредством семантики). Данные выводы могут быть основаны на одном документе или на множестве распределенных документов, которые объединяются с использованием механизмов OWL.

При создании нового класса (например, класса «Вид расходов») ему по умолчанию присваивается стандартное имя (основанное на названии проекта), для изменения которого необходимо изменить значение поля «Name». В системе Protégé приняты правила наименования, согласно которым первая буква в каждом слове в имени класса пишется в верхнем регистре, а остальные буквы – в нижнем. Согласно данным правилам, название для класса «Вид расходов» будет записано как ClassOfExpenditures. В поле Description (описание) сохраняется русское название класса. В системе Protégé классы могут быть как конкретными (Concrete), на основе которых система может создавать экземпляры, так и абстрактными, т.е. классами, у которых экземпляров нет. По умолчанию создается класс «конкретного» типа, изменить который можно в поле Role. Рассматриваемый класс может иметь экземпляры, поэтому тип класса изменять не нужно. Для добавления нового слота необходимо открыть окно создания слота и заполнить все необходимые поля – наименование слота, тип значения, мощность слота и др. У рассматриваемого класса есть два слота – слот codeClassOfExpenditures (код «Вид расходов») и слот nameClassOfExpenditures (наименование «Вид расходов»); в данном случае оба слота имеют единичную мощность и тип данных «строка». Аналогично описанному выше способу создаются все классы из структуры онтологии, представленной на рис. 2.

Для дальнейшего использования полученной онтологии без системы Protégé возможно ее экспортирование в формат RDFs. Фрагмент кода полученного файла представлен в Листинге 1. Тег `<rdfs:Class rdf:about="&rdf_;ClassOfExpenditures">` декларирует некий класс, который описывается в некотором пространстве имен «rdf» как ClassOfExpenditures. Указание различных пространств имен дает возможность повторно использовать другие онтологии, уточнять и расширять их, объявляя новые подклассы. Данная возможность позволяет осуществлять интеграцию узконаправленных информационных ресурсов, объединенных широкой предметной областью.

```
<rdfs:Class rdf:about="&rdf_;ClassOfExpenditures"
   rdfs:comment="Вид расходов"
   rdfs:label="ClassOfExpenditures">
  <rdfs:subClassOf rdf:resource="&rdfs;Resource"/>
</rdfs:Class>
<rdf:Property rdf:about="&rdf_;codeClassOfExpenditures"
   rdfs:label="codeClassOfExpenditures">
  <rdfs:domain rdf:resource="&rdf_;ClassOfExpenditures"/>
  <rdfs:range rdf:resource="&rdfs;Literal"/>
                                  </rdf:Property>
```

Листинг 1. Структура классов в RDFs-файле

Структура кода класса «ClassOfExpenditures» («Вид расходов») заключается в следующем: тег `<rdf:about>` обеспечивает название или ссылку на онтологию, которой в данном случае является ссылка «http://protege.stanford.edu/rdf/ClassOfExpenditures»; тег `<rdfs:comment>` обеспечивает возможность аннотировать онтологию, т.е. добавить комментарий о том, что данный класс обозначает «Вид расходов»; тег `<rdfs:label>` обеспечивает удобное для чтения пользователем название класса, тег `<rdfs:subClassOf>` связывает рассматриваемый класс с его надклассом, которым в данном случае является класс Resource. У рассматриваемого класса есть два свойства: codeClassOfExpenditures и nameClassOfExpenditures, которые в синтаксисе RDFs-файлов прописываются с тегом `<rdf:property>`. Тег `<rdfs:domain>` показывает, что данное свойство относится к свойствам класса ClassOfExpenditures, а тег `<rdfs:range>` обозначает диапазон свойств, ограничивающих экземпляры класса (в данном случае они должны быть типа Literal).

После создания иерархии классов необходимо создать экземпляры, т.е. непосредственно данные базы знаний. Перед этим необходимо еще раз проверить структуру классов, так как при необходимости изменения ее в дальнейшем, возможна потеря уже введенной информации.

Источником данных для класса CEOIncome («Главный распорядитель бюджетных средств») является документ «Перечень и коды главных администраторов доходов бюджета Санкт-Петербурга, которые являются органами государственной власти Санкт-Петербурга, и закрепляемые за ними виды доходов бюджета Санкт-Петербурга». Значениями поля codeCEOIncome являются значения столбца «Код главного администратора бюджетной классификации РФ», значениями поля nameCEOIncome – значения столбца «Наименование».

Источником данных для классов Section («Раздел») и Subsection («Подраздел») являются «Правила отнесения расходов всех бюджетов бюджетной системы РФ на соответствующие классификации расходов»: значением свойства codeSection является номер раздела, nameSection – название раздела, descriptionSection – описание раздела. Аналогично – для класса codeSubsection: свойства codeSubsection – номер подраздела, nameSubsection – название подраздела, descriptionSubsection – описание подраздела, значением свойства belongToSection является экземпляр класса Section, подразделом которого является рассматриваемый экземпляр класса.

Источником значений для классов ClassOfExpenditures («Вид расходов») и TargetArticle («Целевая статья») является Ведомственная структура расходов бюджета Санкт-Петербурга на 2012 год. Для класса ClassOfExpenditures значениями свойства codeClassOfExpenditures являются значения столбца «Код целевой статьи», значения свойства nameClassOfExpenditures – содержание столбца «Наименование» в строках с нумерацией третьего уровня. Аналогично для класса TargetArticle значениями свойства codeTargetArticle являются значения столбца «Код целевой статьи», а значениями свойства nameTargetArticle – содержание столбца «Наименование» в строках с нумерацией второго уровня.

Источником значений для классов OperationOfGovernment («ОСГУ») и Costs («Расходы») является ведомственная структура сводной бюджетной росписи бюджета Санкт-Петербурга на 2012 год. Для класса OperationOfGovernment значениями свойства codeOperationOfGovernment являются значения столбца «Код ОСГУ», для свойства nameOperationOfGovernment – значения столбца «Наименования» для строк с нумерацией третьего уровня. Для класса Costs значениями свойств belongToCEOIncome, belongToClassOfExpenditures, belongToOperationOfGovernment, belongToSubsection, belongToTargetArticle являются соответственно экземпляры классов CEOIncome, ClassOfExpenditures, OperationOfGovernment, Subsection, TargetArticle, у которых значения кодов соответствуют значениям соответствующих столбцов ведомственной структуры сводной бюджетной росписи; значением свойства Sum является значение столбца «Сумма»; значением свойства id является 16-значная строка, состоящая из кодов подраздела, целевой статьи, вида расходов и операции сектора государственного управления.

Создание экземпляров всех классов, кроме экземпляров класса Costs, происходит с помощью интерфейса программы Protégé. После этого полученный проект экспортируется в формат RDFs, получая на выходе два RDFs-файла: файл с иерархией классов (фрагменты файла приводились в Листинге 1) и файл с экземплярами классов (фрагмент файла представлен в Листинге 2). Его структура аналогична структуре файла с иерархией классов, за исключением того, что вместо тегов `<rdfs:Class>` или `<rdf:Proprty>` указывается тег `<rdf:ClassName>`, где ClassName – имя класса, экземпляр которого создается в данном теге.

```
<rdf_:ClassOfExpenditures rdf:about="&rdf_;BudjetSPb2_Class115"
   rdf_:codeClassOfExpenditures="012"
   rdf_:nameClassOfExpenditures="Выполнение функций государственными органами"
   rdfs:label="012"/>
```

Листинг 2. Структура экземпляров в RDFs-файле

Схема автоматизированного создания экземпляров класса Costs представлена на рис. 3. Из первоначального файла ведомственной структуры сводной бюджетной росписи бюджета Санкт-Петербурга на 2012 год (формат PDF) данные копируются в файл формата TXT. После этого с помощью регулярных выражений удаляются все текстовые данные и нумерация строк из столбца «Номер» ведомственной структуры сводной бюджетной росписи. В результате данных действий получен файл, в котором содержатся данные для свойств экземпляра класса Costs: belongToCEOIncome, belongToSubsection, belongToTargetArticle, belongToClassOfExpenditures, belongToOperationOfGovernment и sum, разделенные символом пробела. Данный файл является структурированным, поэтому для его преобразования можно написан парсер на языке программирования Java.

Принцип работы парсера следующий: на входе парсер получает два файла: RDFs-файл с созданными экземплярами класса (source.txt) и полученный txt-файл с данными расходов бюджета Санкт-Петербурга (data.txt). В парсере открываются два потока для чтения данных из файлов и поток для записи результатов

работы программы в файл (result.txt). После считывания всех данных в массив парсер построчно сохраняет значение кодов из файла data.txt в переменные, одноименные будущим свойствам экземпляров класса Costs, и записывает их в выходной файл с соблюдением синтаксиса RDFs-файла. Полученный в результате работы программы файл с экземплярами класса Costs копируется в файл с экземплярами классов онтологии. Таким образом, в результате получается онтология расходов бюджета Санкт-Петербурга на 2012 г., состоящая из двух RDFs-файлов, пригодная для дальнейшей машинной обработки.

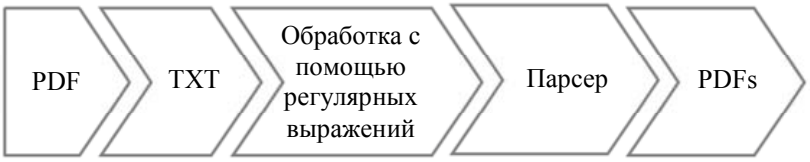

Рис. 3. Схема автоматизированного создания экземпляров класса Costs

Заключение

В работе представлены этапы разработки онтологии ведомственной структуры расходов бюджета Санкт-Петербурга – постановка задачи, исследование структуры данных, проектирование онтологии и ее реализация в системе Protégé. Результатом работы является описание методики создания онтологии открытых государственных данных и разработанный пример в формате RDFs-онтология, состоящий из 7 базовых классов и около 8 тыс. экземпляров. Полученную онтологию можно использовать в качестве источника данных для онлайн-сервисов, позволяющих осуществлять автоматизированную обработку и анализ открытых государственных данных. Примером такого сервиса может являться прототип системы публикации открытых государственных данных расходов бюджета, также полученный в рамках представленной работы. К реализованным функциям прототипа относятся загрузка массива данных (онтологии) и просмотр метаинформации о нем, выбор данных для построения таблиц и графиков и их отображение. На данный момент разработка находится на стадии оптимизации работы прототипа.

Литература

1. Доклад «Об обеспечении доступа населения к информации о деятельности государственных органов и органов местного самоуправления» [Электронный ресурс]. – Режим доступа: http://ivan.begtin.name/wp-content/uploads/2011/09/doklad.pdf, свободный. Яз. рус. (дата обращения: 17.05.2012).
2. Закон Санкт-Петербурга от 26.10.2011 № 658-120 «О бюджете Санкт-Петербурга на 2012 год и на плановый период 2013–2014 годов».
3. Законы о бюджете Санкт-Петербурга // Комитет финансов Санкт-Петербурга [Электронный ресурс]. – Режим доступа: http://www.fincom.spb.ru/comfin/budjet/laws.htm, свободный. Яз. рус. (дата обращения: 01.10.2012).
4. Портал открытого бюджета города Москвы [Электронный ресурс]. – Режим доступа: http://budget.mos.ru, свободный. Яз. рус. (дата обращения: 01.10.2012).
5. Указания о порядке применения бюджетной классификации РФ (утверждены Приказом Министерства финансов РФ от 28.12.2010 № 190н) [Электронная версия]. – Режим доступа: http://www1.minfin.ru/ru/budget/classandaccounting/classification/, свободный. Яз. рус. (дата обращения: 01.10.2012).
6. Муромцев Д.И. Онтологический инжиниринг знаний в системе Protégé. Методическое пособие, Санкт-Петербург, 2007 г. [Электронная версия]. – Режим доступа: http://books.ifmo.ru/book/pdf/243.pdf, свободный. Яз. рус. (дата обращения: 01.10.2012).
7. Сайт ПО Protégé Университет Стэнфорда [Электронный ресурс]. – Режим доступа: http://protege.stanford.edu/, свободный. Яз. рус. (дата обращения: 01.10.2012).
8. OWL Web Ontology Language Guide [Электронная версия]. – Режим доступа: http://www.w3.org/TR/owl-guide/, свободный. Яз. рус. (дата обращения: 01.10.2012.

Пархимович Ольга Владимировна – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, студент, olya.parkhimovich@gmail.com

Власов Виталий Владимирович – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, ассистент, inxaoc@gmail.com

Муромцев Дмитрий Ильич – Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики, кандидат технических наук, доцент, mouromtsev@mail.ifmo.ru