Отчет за 2018 год

по проекту 17-07-00228

МЕтодология и ТЕхнология формирования Онтологий на основе интегРации с гетерогенными источниками данных (МЕТЕОР)

Коды классификатора: **07-966, 07-971**

Ф.И.О. руководителя проекта: Гаврилова Татьяна Альбертовна

Аннотация отчета

Проект МЕТЕОР ориентирован на решение актуальной проблемы проектирования и формирования баз знаний предметных областей на основе онтологий через интеграцию и обработку структурированных массивов данных из гетерогенных источников.

Проект состоит из трех этапов:

1. Анализ состояния современных исследований в области формирования и поддержки баз знаний онтологического типа и разработка основ методологии интеграции структурированных данных из гетерогенных источников в базы знаний различных предметных областей – корпоративных, научных, учебных и др.
2. Создание МЕТЕОР-технологии формирования онтологий и интеграции структурированных данных из гетерогенных источников с учетом человеческого фактора, в частности, когнитивных аспектов дизайна и генерации онтологий.
3. Разработка архитектуры прототипной системы МЕТЕОР-демо, поддерживающей/реализующей эту технологию и апробация разработанных методологии и технологии на примере создания базы знаний учебного процесса ВУЗа.

В 2018 году выполнялся второй этап проекта по созданию новой технологии формирования баз знаний на основе онтологий, основанной на современных подходах к интеграции, учитывающих семантику предметной области (Domain Semantics-Driven Mappings). Этап был посвящен разработке новой технологии, позволяющей компаниям эффективно создавать базы знаний и затем интегрировать в них ранее накопленную информацию и данные. Разработанный новый подход учитывает особенности жизненного цикла онтологий, в частности, проблему устаревания данных. Также предложена типовая архитектура системы.

На втором этапе были решены следующие задачи:

Раздел 1: Создание технологии МЕТЕОР формирования онтологий и интеграции структурированных данных из гетерогенных источников с учетом когнитивных аспектов:

Задача 1-1: формальное описание методологии формирования и наполнения онтологий;

Задача 1-2: создание онтологий источников данных различных типов (реляционная база данных, XML-документ, электронная таблица MS Excel);

Задача 1-3: разработка рекомендаций по описанию источников данных с использованием созданных онтологий.

Раздел 2: Разработка пилотного прототипа программного комплекса, реализующего технологию METEOP, включающая:

Задача 2-1: разработку архитектура системы, поддерживающей/реализующей технологию METEOP;

Задача 2-2: разработку препроцессора, формирующего промежуточную онтологию, содержащую данные из гетерогенных источников;

Задача 2-3: формирование рекомендаций по использованию созданного прототипа.

Все задачи второго этапа проекта METEOP выполнены.

Результаты работы по второму этапу проекта за 2018 год нашли отражение в 13 научных публикациях (7 на английском языке), из них 4 статьи в рецензируемых научных журналах (2 зарубежных журнала, 2 журнала в списке ВАК), и 9 докладов на международных научных конференциях. Публикации проиндексированы: в РИНЦ – 9, в Scopus – 6 и в WoS – 2.

Заявленные цели на 2018 год

Второй этап по созданию новой технологии формирования баз знаний на основе онтологий будет основываться на современных подходах к интеграции, учитывающих семантику предметной области (Domain Semantics-Driven Mappings). Он будет посвящен разработке новой технологии, позволяющей компаниям эффективно создавать базы знаний и затем интегрировать в них ранее накопленную информацию и данные. Подход будет учитывать особенности жизненного цикла онтологий, в частности, проблему устаревания данных. Также будет предложена типовая архитектура системы.

Задачи 2-ого этапа:

1. Создание технологии METEOP формирования онтологий и интеграции структурированных данных из гетерогенных источников с учетом когнитивных аспектов:
    - формальное описание технологии формирования и наполнения онтологий (ГавриловаТ.А., Лещева И.А., Кудрявцев Д.В.)
    - создание онтологий источников данных различных типов (реляционная база данных, XML-документ, электронная таблица MS Excel); (ГавриловаТ.А., Лещева И.А., Беглер А.М., Кубельский М.В.)
    - разработка рекомендаций по описанию источников данных с использованием созданных онтологий; (Плешкова А.Ю., Благов Е.Ю., Страхович Э.В.)
2. Разработка архитектура системы, поддерживающей/реализующей технологию METEOP (Лещева И.А., Кудрявцев Д.В., Кубельский М.В.).
3. Разработка пилотного прототипа программного комплекса, реализующего технологию METEOP, включающая:
    - разработку препроцессора, формирующего промежуточную онтологию, содержащую данные из гетерогенных источников; (Лещева И.А.)
    - формирование рекомендаций по использованию созданного прототипа (Беглер А.М.).

Полученные за 2018 важнейшие научные результаты

Второй этап был посвящен созданию технологии формирования онтологий и интеграции структурированных данных из гетерогенных источников с учетом когнитивных аспектов, а также разработке архитектуры системы, поддерживающей/реализующей эту технологию. Под источниками данных понимаются реляционные БД, электронные таблицы и т.п.

Результаты по задачам раздела 1

Задача 1-1: формальное описание технологии формирования и наполнения онтологий

Методология METEOP для формирования онтологий на основе интеграции с гетерогенными источниками данных является обобщением существующих методов консолидации данных и наполнения онтологий, учитывающих семантику предметной области [Petasis et al., 2011]. Для этого используется описание отображений в онтологической форме, что позволяет расширить область применения подходов [Pan, 2017] на консолидацию данных, представленных в табличном и XML-формате.

В рамках проекта предложен новый метод наполнения онтологий, который состоит из 4 шагов (см. рис. 1):

1) **Создание онтологии**. На этом этапе инженер по знаниям совместно с экспертами предметной области формируют начальную онтологию и формулируют цель онтологии. При разработке начальной онтологии могут быть использованы как любые методологии, так и любые универсальные или расширенные редакторы онтологий (например, приведенные в отчете за 2017 г.). При выборе инструмента для работы с онтологиями рекомендуется учитывать особенности организации и класс задач, для решения которых планируется использовать онтологии [Беглер, Костоусов, 2018]. Необходимо отметить, что практически все методологии включают шаг, предполагающий определение области и масштаба онтологии.
2) **Определение источников данных**. Этот этап включает аудит на предмет выявления источников данных, необходимых для решения задач согласно цели онтологии. В качестве источников могу выступать базы данных, электронные таблицы, XML-файлы и другие файлы проприетарного формата, в которых хранятся структурированные данные.
3) **Формирование спецификации источников данных**. Особенностью предлагаемого метода является отсутствие специализированного языка описания отображений. Вместо этого предлагается специфицировать источники данных и описывать отображение с помощью мета-шаблона, который представляет собой легковесную онтологию, описанную на RDF, и задает структуру для всех свойств источников данных, а также для описания особенностей отображения. Подробнее мета-шаблон описан в пункте *Задача 1-2,* а процесс формирования спецификации источников данных в пункте *Задача 1-3.*
4) **Наполнение онтологии.** Используя созданную на предыдущем шаге спецификацию, разработанная на первом шаге онтология наполняется данными из источников, выбранных на втором шаге. Подробнее процесс наполнения описан в пунктах *Задача 2-1* и *Задача 2-2.*

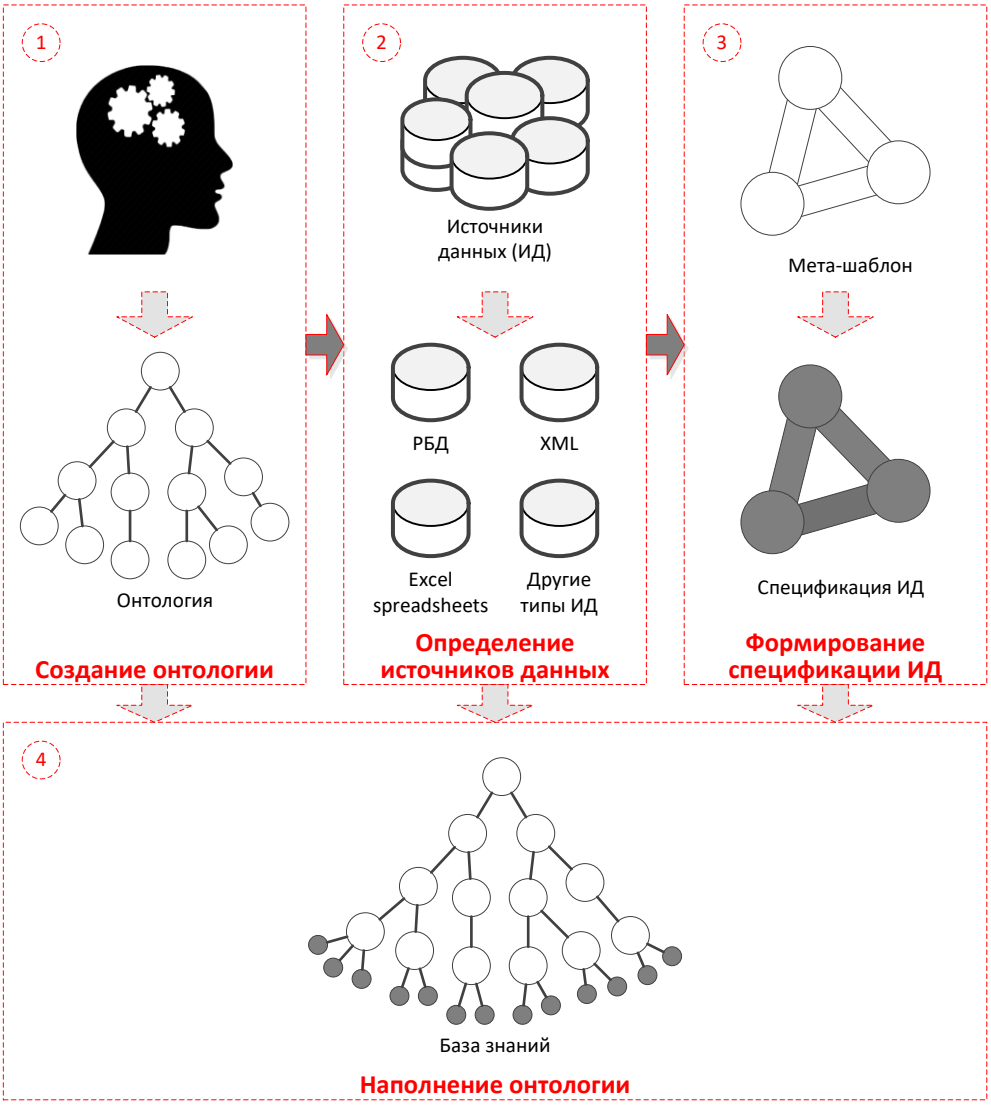

Рис. 1. *Метод наполнения онтологий*

В отличие от существующих методов, предполагающих однократное наполнение онтологии, предлагаемый метод можно использовать для односторонней синхронизации данных (причем это не требует повторения шагов 1-3). Также этот метод делает возможным одновременное наполнение онтологии данными из источников различных типов (с учетом ее семантики), при этом не требует изучения синтаксиса специфического языка описания отображений и/или разработки инструментария для облегчения создания описания отображения.

Задача 1-2: создание онтологий источников данных различных типов

В рамках выполнения проекта была создана онтология источников данных (мета-шаблон для последующего описания свойств источников данных). Данная онтология основана на спецификациях этих источников и на онтологиях, найденных в открытом доступе. Ниже описаны основные классы такой онтологии. Класс **DataSource** — это класс верхнего уровня, объединяющий разные типы источников данных. Любой источник данных имеет название, путь, а также могут потребоваться аутентификационные данные для получения доступа к этому источнику. Класс **DataSource** включает такие подклассы как **Workbook** (для табличных источников данных), **Database** (для реляционных баз данных) и **XMLdocument** (для xml-документов). На рисунке 2 отражены все

существенные классы и отношения мета-шаблона. Для визуализации использовалась нотация из книги «Demystifying OWL for the Enterprise» [Uschold, 2018].

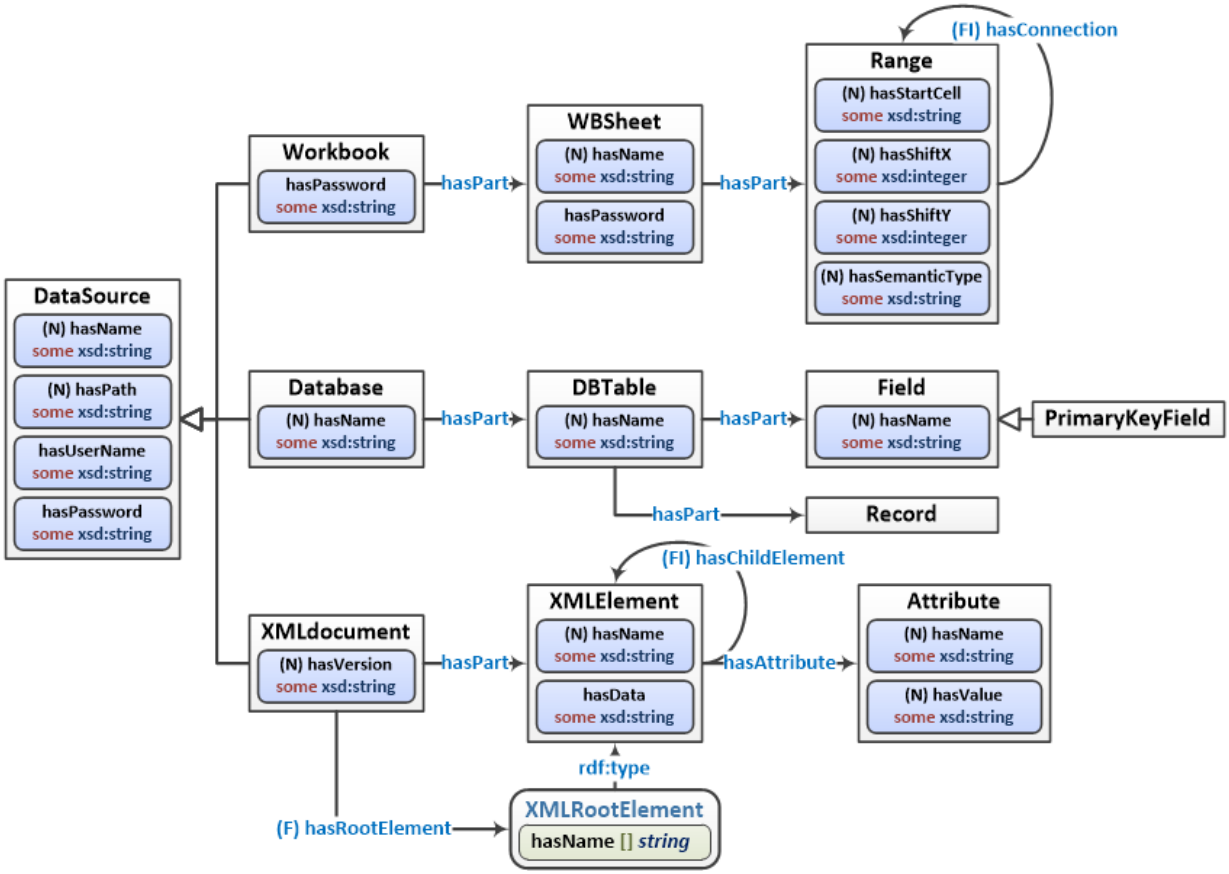

Рис. 2. *Онтология источников данных — мета-шаблон*

Класс **Workbook,** предназначенный для описания табличных данных, состоит из экземпляров класса **WBSheet**, описывающих лист таблицы. Диапазоны ячеек являются частью листа и описываются с помощью класса **Range**.

Класс **Range** (диапазон) — это ключевой класс для описания табличных данных. Именно он связывает значения в ячейках таблицы с онтологией. Каждая выделенная группа ячеек становится диапазоном, который описывается рядом свойств. Такой подход был реализован для преодоления популярного “entity-per-row assumption”, при котором подразумевается, что в каждой строке таблицы описывается одна сущность, а в столбцах — свойства этой сущности. [O’Connor, Halaschek-Wiener, Musen, 2010] отмечают, что таблицы в реальном мире зачастую не соответствуют этому предположению и могут содержать в столбцах, например, изменения характеристик сущности в разные моменты времени (например, в финансовых таблицах в столбцах могут содержаться месяца). В таких случаях каждая клетка таблицы становится сущностью (entity) в OWL, которая может быть как классом, так и экземпляром, и обладать свойствами. Описываемый подход также позволяет описывать таблицы произвольной структуры за счёт того, что каждый диапазон ячеек (экземпляр класса Range) характеризуется рядом свойств (описание этих свойств на даётся в отчёте из соображений краткости). В пункте *Задача 1-3* приводится пример описания таблицы с помощью предложенного мета-шаблона.

В литературе предложен целый ряд подходов к интеграции данных в онтологию из реляционных баз данных [Michel, Montagnat, Faron-Zucker, 2014]. Описываемый подход отличается тем, что для спецификации отображения используется не специальный язык, а визуализированное описание на основе мета-шаблона, представленного в формате RDF. Для заполнения этого шаблона нет необходимости изучать синтаксис языков отображения или использовать специализированные инструменты — подойдет любой редактор онтологий.

Любая реляционная база данных (класс **Database**) содержит одну или несколько таблиц (класс **DBTable**), которые, в свою очередь, состоят из полей (**Field**) и содержат записи (**Record**). Некоторые поля могу быть первичными ключами, т.е. являться экземплярами подкласса **PrimaryKeyField**. Свойства записей формируются на 4 шаге и их формирование описывается в *Задаче 2-2*.

XML (англ. eXtensible Markup Language) — расширяемый язык разметки, рекомендованный Консорциумом Всемирной паутины (W3C). Основным документом, описывающим данный язык является англоязычная спецификация, представленная на сайте Консорциума (https://www.w3.org/TR/xml/). На сегодняшний день XML является наиболее часто используемым стандартом для представления, структурирования и обмена данными в интернете. Таким образом, данные, представленные в виде XML-документов могут являться ценными источниками знаний как для развития схемы онтологии (TBox), так и для обогащения существующих схем экземплярами (ABox).

Помимо представленных на рисунке 2, используются вспомогательные отношения. Главное из них — отношение ***correspondsTo***, с помощью которого можно показать, что данные, содержащиеся в источниках различного типа, относятся к одним и тем же объектам, а также определить, на основании значений каких свойств можно идентифицировать одинаковые объекты.

Задача 1-3: разработка рекомендаций по описанию источников данных с использованием созданных онтологий

Так как и реляционная база данных, и XML-документ содержат в себе свою структуру (т.е. получив доступ к базе данных, можно узнать имена таблиц, их полей и т.п.; также можно узнать все об элементах XML-документа), то необходимо только задать значения свойств, указанных в описании классов **DataSource** и **Database**, и первичное формирование спецификации этих источников данных происходит автоматически. Затем с помощью вспомогательных отношений (например, упомянутого выше ***correspondsTo***) настраивается отображение в онтологию.

Для трансляции таблиц в онтологию необходимо: 1) разметить все имеющиеся ячейки с данными на листе как диапазоны; 2) назначить диапазонам имена; 3) описать каждый диапазон с использованием описанных выше свойств.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Сведения об успеваемости обучающихся на основных образовательных программах бакалавриата												
2													
3	Программа:		Менеджмент										
4	Год поступления:		2016										
5													
6				1 семестр					2 семестр				
7	№	ФИО	Профиль	История бизнеса	Правоведение	Макроэкономика	Математика	Средний балл	Микроэкономика	Статистика	Менеджмент	Философия	Средний балл
8	1	Андреев Иван Петрович	ФМ	4	3	4	4	3,8	3	3	4	4	3,6
9	2	Бондарь Наталья	ИМ	5	4	5	4	4,5	4	4	5	5	4,5
10	3	Вавилов Сергей Евгеньевич	УЧР	3	4	5	5	4,3	5	5	4	3	4,3

Рис. 3. *Пример таблицы*

В качестве примера опишем с помощью описанного выше мета-шаблона таблицу, показанную на рисунке 3. Первым шагом описания будет разметка диапазонов ячеек в таблице (рис. 4).

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Сведения об успеваемости обучающихся на основных образовательных программах бакалавриата												
2													
3	Программа: (1)		Менеджмент (2)										
4	Год поступления: (3)		2016 (4)										
5													
6				1 семестр (9)					2 семестр (14)				
7	№	(5) ФИО	(7) Профиль	История бизнеса	Правоведение	Макроэкономика	(10) Математика	(12) Средний балл	Микроэкономика	Статистика	Менеджмент	(15) Философия	(17) Средний балл
8	1	Андреев Иван Петрович (6)	(8) ФМ	4	3	4	(11)	(13)	3	3	4	(16)	(18)
9	2	Бондарь Наталья	ИМ	5	4	5	4	4,5	4	4	5	5	4,5
10	3	Вавилов Сергей Евгеньевич	УЧР	3	4	5	5	4,3	5	5	4	3	4,3

Рис. 4. *Разметка диапазонов ячеек таблицы*

Далее озаглавим диапазоны и опишем их. Ниже приведено описание диапазона 11.

range 11

hasStartCell – D8

hasShiftX – 3

hasShiftY – 2

hasSemanticType – instance

Также диапазон **range11** связан отношением ***hasConnection*** с диапазонами **range6** и **range10**.

Значение свойства ***hasSemanticType*** зависит от структуры онтологии, для наполнения которой предполагается использовать таблицу. Оно может принимать следующие значения: className, instance, dataPropertyName, dataPropertyValue.

В таблице 1 указаны возможные комбинации значений свойства ***hasSemanticType*** у диапазонов, связанных отношением ***hasConnection***.

Таблица 1. Возможные комбинации значений свойства ***hasSemanticType*** у диапазонов, связанных отношением ***hasConnection***

A \ **B**	**className**	**instance**	**dataPropertyName**	**dataPropertyValue**
className	Вариант1	-	-	-
instance	Вариант2	Вариант4	-	-
dataPropertyName	Вариант3	-	-	-
dataPropertyValue	-	Вариант5	Вариант6	-

Далее при описании семантики вариантов, указанных в таблице 1, предполагается, что 1) диапазон **A** связан отношением ***hasConnection*** с диапазоном **B**; 2) значение свойства ***hasSemanticType*** для диапазона **A** берется из соответствующей ячейки первого столбца; 3) значение свойства ***hasSemanticType*** для диапазона **B** берется из соответствующей ячейки первой строки.

Вариант 1. В ячейках диапазона **A** описываются имена подклассов классов, имена которых указаны в ячейках диапазона B.

Вариант 2. Ячейки диапазона **A** соответствуют экземплярам классов, имена которых указаны в ячейках диапазона B.

Вариант 3. Ячейки диапазона **A** соответствуют именам свойств данных (dataProperties), присущих экземплярам классов, имена которых указаны в ячейках диапазона **B**.

Вариант 4. Показывает, что экземпляры, соответствующие диапазону **A**, связаны отношением ***relatesTo*** с экземплярами, соответствующими диапазону **B**.

Вариант 5. В ячейках диапазона **A** содержатся значения свойств экземпляров, соответствующих диапазону **B**.

Вариант 6. В ячейках диапазона **A** содержатся значения свойств, имена которых указаны в ячейках диапазона **B**.

Две последних комбинации должны работать в паре, то есть диапазон A должен быть связан отношением ***hasConnection*** с двумя другими диапазонами.

Результаты по задачам раздела 2

Задача 2-1: разработка архитектура системы, поддерживающей/реализующей технологию METEOP

Архитектура системы, реализующей технологию METEOP, сходна с архитектурой ETL-систем [Паклин, Орешков, 2013], так как они выполняют аналогичные задачи (см. рис. 5). Основное отличие заключается в том, что получателем данных является не хранилище или база данных, а онтология или база знаний, что обуславливает и остальные особенности.

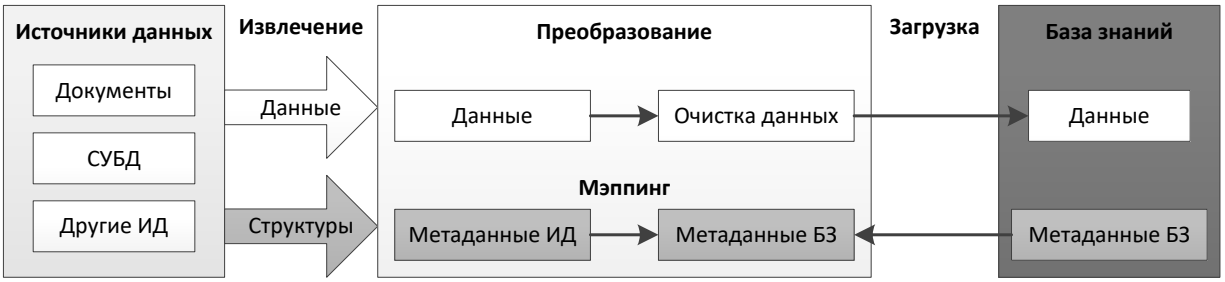

Рис. 5. *Обобщенная структура наполнения базы знаний*

С точки зрения процесса наполнения онтологии архитектуру системы можно представить в виде пяти компонентов (рис. 6), таких как:

- источники данных — содержит структурированные данные в виде совокупности таблиц и/или файлов, данные в которых, например, упорядочены в типизированные столбцы, отделенные друг от друга некоторыми символами-разделителями;
- промежуточная область — содержит вспомогательный граф знаний — набор триплетов, базирующийся на мета-шаблонах — создаваемый временно и исключительно для организации процесса наполнения;
- получатель данных — онтология, в которую должны быть помещены извлеченные данные;
- правила обработки — требования к организации потоков данных описываются аналитиком с помощью мета-шаблонов (см. *Задачу 1-2*);
- редактор онтологий — используется в качестве инструмента при создании онтологии и при описании правил обработки или отображения.

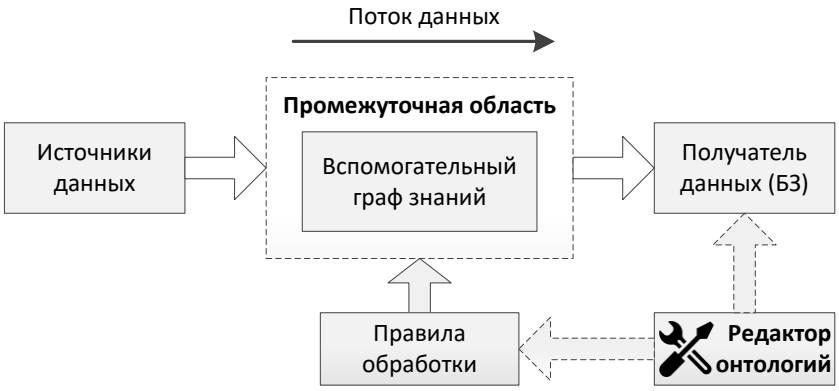

Рис. 6. *Поток данных при наполнении онтологии*

Особенностью подхода является то, что правила обработки задаются аналитиком при взаимодействии с экспертом предметной области в виде онтологии, что не требует специальных инструментов, навыков программирования или знания специализированных языков описания отображений.

Задача 2-2: разработка препроцессора, формирующего промежуточную онтологию, содержащую данные из гетерогенных источников

Прототип состоит из следующих модулей:

1. Модуль анализа спецификации источников данных: загружает и анализирует информацию из спецификации ИД, создает в памяти шаблон вспомогательного графа знаний.
2. Модуль извлечения данных: получает доступ к источникам данных и загружает все данные во вспомогательный граф знаний, при этом сохраняется информация о том, откуда именно был загружен каждый элемент данных.
3. Модуль мэппига: устанавливает соответствие между сущностями из вспомогательного графа знаний и наполняемой онтологией, используя правила, заданные аналитиком при помощи эксперта.
4. Модуль очистки данных: обнаруживает и устраняет противоречия в данных, используя механизм приоритетов, и формирует отчет о найденных ошибках.
5. Модуль загрузки: данные преобразуются к виду, соответствующему структуре наполняемой онтологии и загружаются в нее.

Продемонстрируем работу прототипа для онтологии из области образования, состоящей из 3 классов: **student** (студент), **discipline** (дисциплина), **mark** (оценка) и определяющей 3 свойства (ObjectProperty): ***relatesTo*** [F] (domain: mark, range: discipline); ***receivedBy*** [F] (domain: mark, range: student); ***attends*** (domain: student, range: discipline). [F] означает, что свойство функциональное. На рисунке 7 представлена концептуальная схема онтологии.

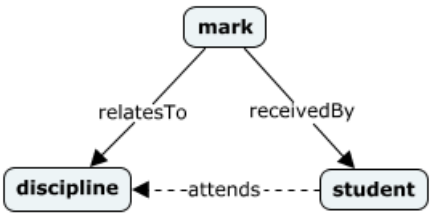

Рис. 7. *Концептуальная схема онтологии*

Данные о дисциплинах, студентах и их оценках содержатся в базе данных. Упрощенная схема фрагмента базы данных показана на рисунке 8.

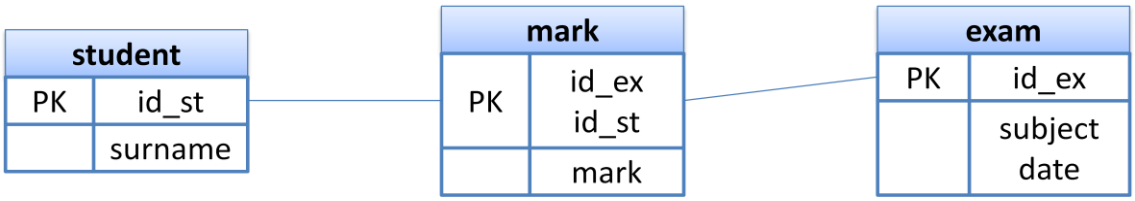

Рис. 8. *Схема фрагмента базы данных*

Тогда результатом работы 1-го модуля является структура, показанная на рисунке 9.

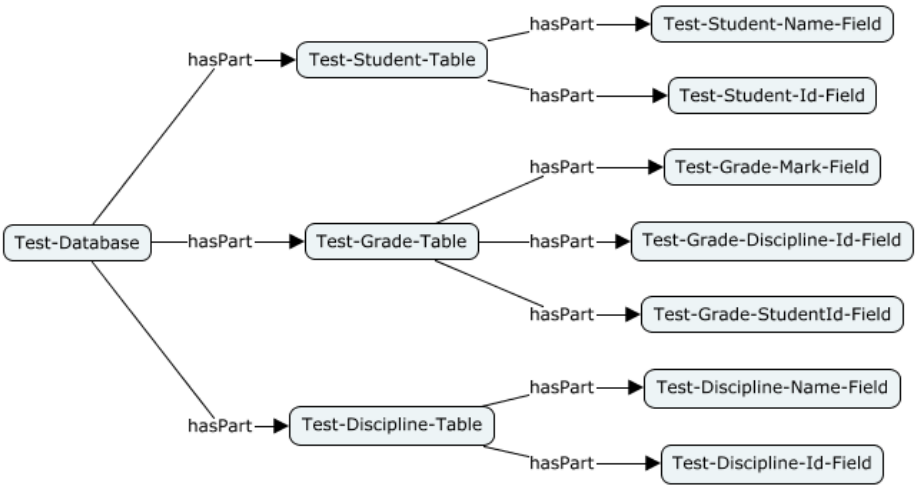

Рис. 9. *Результат работы 1-го модуля*

На рисунке 10 — результат работы 2-го модуля. Данные «встраиваются» в структуру как значения свойств соответствующих записей (экземпляры, чье имя оканчивается на «Record»), поэтому на рисунке не видны.

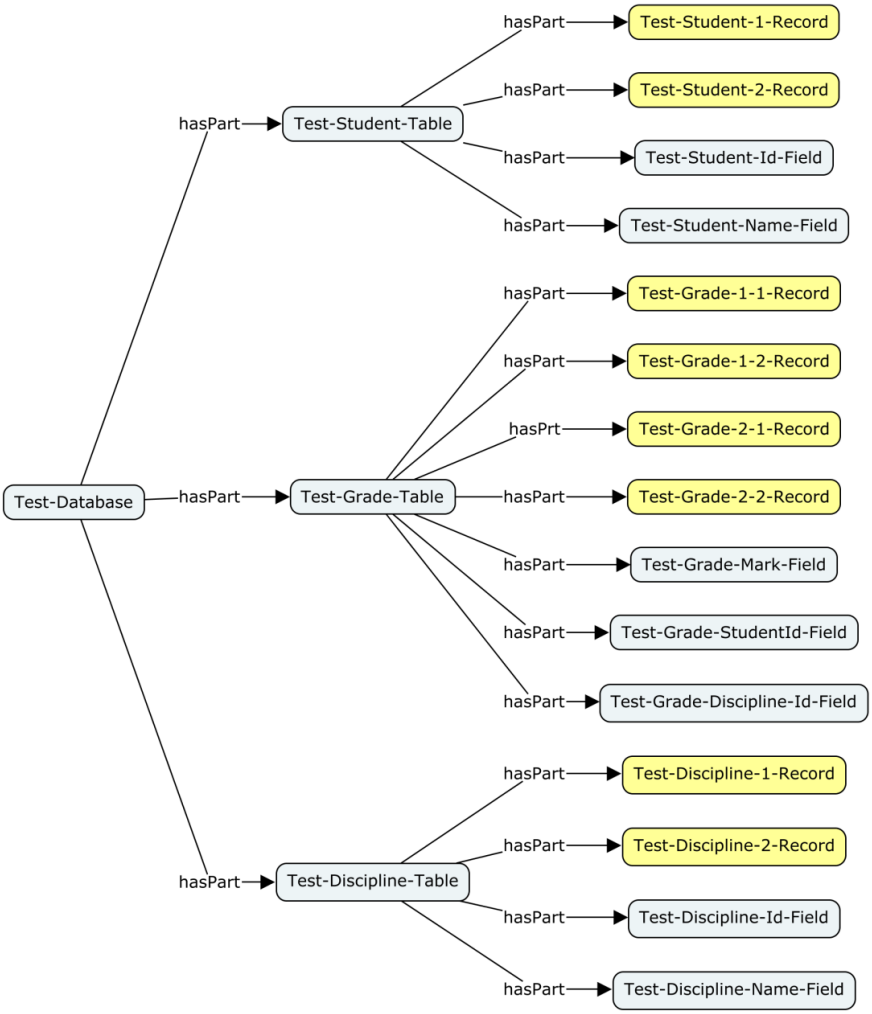

Рис. 10. *Результат работы 1-го модуля*

В результате работы 3-го модуля фиксируется, что записи, содержащиеся в Test-Discipline-Table соответствуют экземплярам класса discipline и т.д.

Модуль очистки данных в данном случае ошибок не обнаружит. Но на практике возможна ситуация, что оценки, полученные из разных источников, не совпадают между собой. Тогда предпочтение отдается оценке, полученной из наиболее приоритетного источника, а информация о найденном несоответствии заносится в файл отчета.

Далее данные загружаются из вспомогательного графа в онтологию. Результат (для наглядности только для 2 студентов и 2 дисциплин) показан на рисунке 11. В данном случае названия дисциплин, значения оценок и т.п. являются значениями соответствующих свойств, поэтому не видны при визуализации в графовой форме.

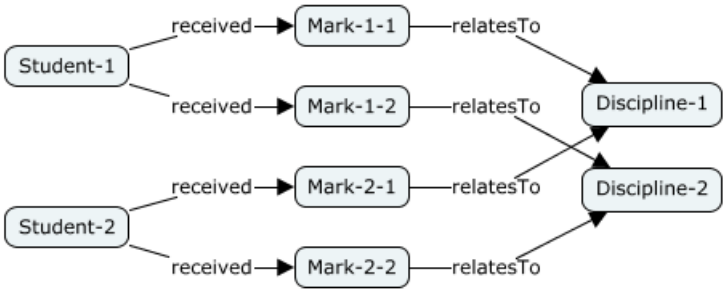

Рис. 11. *Результат наполнения онтологии*

Задача 2-3: формирование рекомендаций по использованию созданного прототипа

Предлагаемая методика направлена на осуществление однотипных операций или на решение повторяющихся задач, требующих знаний правил и процедур, а также сопоставления информации из различных источников. Такого рода задачи возникают на предприятиях, в учебных заведениях, а также в организациях различной ведомственной принадлежности в рамках цифрового документооборота. Предложенная методика может быть положена в основу систем поддержки принятия решений, основанных на знаниях.

Например, в системах электронной коммерции широкое распространение получили рекомендующие системы. Используются различные подходы к составлению списка рекомендаций: коллаборативные рекомендации, рекомендации на основе содержимого, рекомендации на основе знаний. Последний подход может быть реализован, в частности, с помощью предложенного метода. В онтологию закладываются знания о критериях выбора, совместимости объектов и т.п. Характеристики имеющихся объектов могут храниться в базе данных или отдельных файлах. Информация о предпочтениях пользователя может содержаться в запросе. По запросу данные об объектах «втягиваются» в онтологию, анализируются в соответствии с заложенными правилами и на основе проведенного анализа формируется рекомендация для конкретного пользователя.

Разработанный прототип реализует созданную методику и позволяет практически создавать и наполнять базы знаний онтологического типа из гетерогенных источников информации.

Результаты работы по второму этапу проекта за 2018 год нашли отражение в 13 научных публикациях (7 на английском языке), из них 4 статьи в рецензируемых научных журналах (2 зарубежных журнала, 2 журнала в списке ВАК), и 9 докладов на международных научных конференциях. Публикации проиндексированы: в РИНЦ – 9, в Scopus – 6 и в WoS – 2.

Степень новизны полученных результатов

Новизну разработанных в рамках второго этапа проекта METEOP методов и полученных результатов можно подразделить на новизну по отношению к мировому контексту исследований в области онтологического моделирования корпоративных знаний в целом и по отношению к российскому контексту подобных исследований.

Относительно новизны в мировом контексте следует отметить ориентацию разрабатываемой методологии на преодоление таких недостатков существующих подходов к наполнению и обогащению онтологий как [Valarakos et al., 2004; Petasis et al., 2011; Nederstigt et al., 2014]: узкая направленность на решение специфических задач конкретных проектов, обеспечение возможности интеграции данных только из однотипных источников, отсутствие модульности и расширяемости и недостаток удобного инструментария. Предлагаемая методология позволяет использовать для интеграции данные различных типов (в частности, реляционные базы данных, XML-документы, электронные таблицы MS Excel) и таким образом создавать и модифицировать, в т.ч., расширять, онтологии различных предметных областей.

Новизна разрабатываемой методологии в российском контексте заключается в самом фокусе на интеграцию разнородных данных в онтологии предметной области с учётом когнитивных аспектов дизайна и генерации онтологий. Большинство российских работ, посвящённых интеграции данных с использованием онтологий (напр., работы коллективов под руководством В.Ф. Тузовского, М.Р. Когаловского и др.), используют онтологии в качестве инструмента, но не предполагают создания новых, либо модификации существующих онтологий в качестве результата работы. Таким образом, предлагаемая в рамках проекта METEOP методология в контексте российских исследований схожей тематики обладает высокой степенью новизны и актуальности как с теоретической и методологической точки зрения, так и с точки зрения практической реализации разработанного пилотного прототипа, позволяющего решать задачи, до этого не решавшиеся программными средствами российских разработчиков.

Представляется, что вышеописанные отличия предлагаемой методологии от российских и зарубежных аналогов делают данную методологию достаточно перспективной с точки зрения внедрения в практику деятельности организаций различных сфер деятельности, в частности, к примеру, образовательных учреждений [Благов, Лещёва, Щербан, 2018], эффективность и результативность деятельности которых в особенной степени зависит от способности максимально логичным и непротиворечивым образом структурировать используемые в работе данные, информацию и знания различного характера.

Сопоставление полученных результатов с мировым уровнем

Проблематика наполнения и обогащения онтологий является достаточно популярной областью исследований. На настоящий момент это одно из наиболее как теоретически и методологически глубоких, так и непосредственно практически ориентированных направлений управления знаниями. Однако, как указано выше, существующие методологии отличаются целым рядом недостатков, придающих технологии проекта METEOP существенную степень уникальности.

Начиная с 1998 года, когда появилась первая версия XML, было реализовано множество разработок и исследований, связанных с преобразованием XML-файлов в

формальные онтологии, представленные в форматах OWL и RDF [Erdmann and Studer, 2001; Quix et al., 2007; Thuy et al., 2008; Aussenac-Gilles and Kamel, 2009; Bakkas et al., 2014]. Согласно [Hacherouf et al., 2015] существующие подходы можно разделить на две основные группы. Первая группа, это подходы, ориентированные на экземпляры (instance approaches), которые позволяют работать с произвольными XML-документами без наличия соответствующей XML-схемы [JXML2OWL, XMLMaster]. Ко второй группе относятся валидационные подходы (validation approaches), основанные на обработке XML-схем, используемых для валидации соответствующих XML-документов.

Данные табличного типа (tabular form) широко распространены и используются рядом приложений: текстовыми процессорами (text processor), электронными таблицами (spreadsheet), программами для подготовки презентаций (presentation program) и языками разметки (markup language). Большинство существующих решений для интеграции данных рассматривают электронные таблицы и языки разметки, в частности HTML. Таблицы включают в себя как данные (эксплицитно), так и их схему (имплицитно, в структуре строк и столбцов). Таким образом, из таблиц можно извлекать как данные (ABox) для наполнения онтологии, так и непосредственно структуру онтологии (TBox) для её обогащения, а также свойства (RBox) [Compton et al., 2015]. Существующие решения охватывают как извлечение схемы источника данных (например, [Tijerino et al. 2005] для HTML-таблиц), так и извлечение данных в ранее созданную схему. Принципиальная разница в интеграции данных этих двух подходов в том, откуда берётся структура получившейся базы данных – в первом случае этого структура источника данных (direct mapping), во втором она создаётся заранее (domain semantic-driven mapping).

С технической точки зрения разрабатываемая технология является расширением и дополнением существующих методов консолидации данных и наполнения онтологий, учитывающих семантику предметной области. В частности, в отличие от методов, рассматриваемых и предлагаемых в [Бова, 2015; Chungoora et al., 2013; Sitharamulu, Babu, 2016; Paulheim, 2017], технология проекта METEOP позволяет производить не только однократное наполнение онтологии, но и синхронизацию данных и наполнение онтологии данными различных форматов и источников.

К более специфичным чертам разрабатываемой методологии (впрочем, непосредственно связанным с её общей идеологией), отличающим её от мирового уровня, можно отнести то, что при формировании спецификации источников данных для наполнения онтологий не предполагается формирование специализированного языка описания отображений, вместо которого используется мета-шаблон, представляющий собой легковесную онтологию; таким образом, разрабатываемая методология приобретает дополнительную гибкость.

Соответственно, можно заключить, что разрабатываемая в рамках проекта METEOP методология, находясь в русле ключевых и наиболее динамично развивающихся направлений исследований в области управления знаниями в целом и построения и поддержки онтологий в частности, отличается рядом черт общей архитектуры и конкретных решений. Эти черты позволяют максимально гибко интегрировать в онтологии данные различных типов с учётом проблем устаревания данных и когнитивных аспектов дизайна и генерации онтологий. Данная гибкость придаёт результатам проекта METEOP, в частности, его второго этапа, потенциально высокую степень практической

применимости для управления знаниями организаций различных отраслей и бизнес-моделей. Представляется, что помимо образовательной отрасли, особенную актуальность проблематика объединения в рамках единой онтологии разнородных данных имеет в российском контексте в отношении производственных отраслей – вследствие кадрового кризиса в период 1990-х гг. эти области испытывают затруднения с межвременным сохранением и передачей технологических и организационных знаний [Бодрова, Голованова, 2017; Балацкий, Юревич, 2018]. Таким образом, результаты работы, помимо теоретической и методологической, имеют и безусловную практическую значимость для различных отраслевых контекстов.

Ценность полученных на втором этапе результатов подтверждается и успешным представлением данных результатов на различных как российских, так и международных научных и научно-практических конференциях.

Методы и подходы, использованные в ходе выполнения проекта

Разработка нового метода наполнения онтологий и интеграции структурированных данных из гетерогенных источников была основана на системном подходе и на существующих технологиях интеграции, учитывающих семантику предметной области (Domain Semantics-Driven Mappings), использующих языки описания отображений (mapping description languages).

Проект Метеор в целом выполняется в парадигме «проектирующего исследования» (design science research), в соответствие с которой исследователь решает актуальные проблемы путем создания инновационных артефактов, пригодных для повторного использования и вносящих вклад в научное знание [Hevner, Chatterjee, 2010].

«Проектирующее исследование» в соответствие с [Österle et al, 2011] должно следовать 4-м основным принципам:

1. Обобщенность – создаваемый артефакт применим к классу проблем;
2. Оригинальность – создаваемый артефакт должен расширять существующие знания и предмете;
3. Обоснованность – создаваемый артефакт должен быть подкреплен понятными аргументами и предоставлять возможность валидации;
4. Полезность – создаваемый артефакт должен приносить пользу – либо в ближайшее время, либо в будущем – для соответствующих заинтересованных сторон.

В соответствие с [Peffers et al, 2007] «проектирующее исследование» включает 6 шагов:

1. Идентификация и обоснование проблемы (актуальность);
2. Идентификация целей, требований и ограничений;
3. Проектирование и разработка артефакта;
4. Демонстрация использования (апробация) созданного артефакта;
5. Оценка созданного артефакта с точки зрения его результативности, эффективности и т.п.;
6. Обнародование полученных результатов.

Данный этап проекта соответствовал третьему шагу (проектирование и разработка артефакта — метода и технологии). Первые два шага были выполнены на предыдущем этапе проекта. 4-й, 5-й и 6-й шаги — предмет следующего этапа проекта.

В целом задачи второго этапа направлены на широкое аналитическое исследование и разработку новых алгоритмов в области онтологического инжиниринга. Для проведения этих работ использовались следующие методы:

1. контекстный анализ;
2. системный анализ;
3. критический структурный обзор литературы, технологий, платформ и сервисов (более 20 источников, список приложен в файле pdf);
4. структурный дизайн интеллект-карт;
5. инструментарий онтологического инжиниринга;
6. инструменты Google Analytics;
7. генерализация, категоризация и синтез понятий для построения таксономий.

В сочетании с «проектирующим исследованием» эти методы использованы впервые, а их синергетический эффект позволил получить новые оригинальные результаты, опубликованные в 13 научных работах.

План работ на 2019 год

Третий этап: Апробация разработанной методологии и технологии на примере создания базы знаний учебного процесса ВУЗа.

Третий этап предполагает дальнейшее исследование по созданию прототипа базы знаний на примере учебного процесса ВУЗа, демонстрирующего преимущества разработанной на предыдущих этапах методологии и архитектуры. База знаний ориентирована на поддержку процессов принятия решений в задачах административного управления основными образовательными процессами.

Задачи 3-го этапа:

1. Анализ административных задач, возникающих в рамках управления и обеспечения учебных процессов. Выбор задачи для кастомизации методологии создания базы знаний и ее архитектуры. (Благов Е.Ю., Лещева И.А., Кубельский М.В., Плешкова А.Ю., Гринберг Э.Я.)
2. Апробация разработанного подхода (Гаврилова Т.А., Лещева И.А., Беглер А.М., Благов Е.Ю., Кудрявцев Д.В., Кубельский М.В.):
    - Формирование пилотного прототипа базы знаний для решения выбранной задачи.
    - Выбор источников данных для заполнения базы знаний.
    - Создание спецификации источников данных.
    - Наполнение базы знаний.
3. Верификация и оценка качества полученной базы знаний. (Кудрявцев Д.В., Лещева И.А., Благов Е.Ю., Беглер А.М.)

Библиографический список всех публикаций по проекту за 2018 год

Статьи

1. Благов Е.Ю., Лещёва И.А., Щербан С.А. Онтологический подход в практике образовательной деятельности: формирование траекторий индивидуального профессионального развития студентов // Открытое образование. – 2018. – №5. – С. 26-39. (журнал индексируется в РИНЦ, список ВАК)
2. Гринберг Э.Я., Плешкова А.Ю. К вопросу о подготовке специалистов по управлению знаниями // Открытое образование. –ч 2018. – №22(2). – С. 14-26. DOI:10.21686/1818-4243-2018-2-14-26 (уже проиндексировано в РИНЦ, список ВАК)
3. Gavrilova T., Alsufyev A., Pleshkova A. Formalizing company KM portrait: pilot study with evidence from Russia // Measuring Business Excellence – 2018. – Vol. 22, No. 3. – P. 315-332. https://doi.org/10.1108/MBE-09-2017-0067 (проиндексировано в РИНЦ, Scopus и WoS)
4. Gavrilova T., Kubelskiy M., Kudryavtsev D., Grinberg E. Modeling methods for strategy formulation in a turbulent environment // Strategic Change (Wiley publishing) – 2018. – Vol. 27, No. 4. – P. 369–377. DOI: 10.1002/jsc.2209 (проиндексировано в РИНЦ, Scopus и WoS)

Доклады

5. Беглер А. М., Костоусов, С. А. Критерии выбора инструмента по работе с корпоративными базами знаний // Сборник студенческих работ V Международной научно-практической конференции “Управленческие науки в современном мире” 2018 / г. Москва, (2018). – С. 18–23. (РИНЦ)
6. Кудрявцев Д.В., Гаврилова Т.А., Лещева И.А., Беглер А., Кубельский М., Тушканова О. Методика групповой работы по визуальной разработке графа знаний // Труды национальной конф. по искусственному интеллекту КИИ 2018 / г. Москва, (2018). – Т.1. – С. 53-61. (РИНЦ)
7. Кудрявцев Д.В., Беглер А.М. Системы организации знаний: виды и примеры использования // Сборник научных трудов XXI Российской конференции «Инжиниринг предприятий и управление знаниями», М., ФГБОУ ВО «РЭУ им. Г.В. Плеханова», 2018. – Т.1. – С. 184-190. (РИНЦ)
8. Лещева И.А. Практический подход к созданию и наполнению баз знаний онтологического типа: проблемы и вызовы. // Сборник научных трудов XXI Российской конференции «Инжиниринг предприятий и управление знаниями», М., ФГБОУ ВО «РЭУ им. Г.В. Плеханова», 2018. – Т.1. – С. 194-201. (РИНЦ)
9. Gavrilova T., Kubelsky M. Knowledge Management Specification: Building An Ontology To Get A Shared Understanding // Abstracts and Conference Materials for the 19th European Conference on Knowledge management ECKM / University of Padua, Italy, (2018). – P.131. Online proceedings URL: https://tinyurl.com/ECKM2018 – P. 1117-1123. (Scopus)
10. Kudryavtsev D., Gavrilova T., Kuznetsova A. The Efficacy of Knowledge Management Tools across Domains // Book of abstracts, 13th International Forum on Knowledge Asset Dynamics IFKAD 2018 / Delft, (2018). – p. 141. Online proceedings URL:

https://www.ifkad.org/wp/wp-content/uploads/documents/papers/y2018/135_IFKAD2018.pdf P. 1785-1794. (Scopus)

11. Kudryavtsev D., Gavrilova T., Leshcheva I., Begler A., Tushkanova O. Mind Mapping and Spreadsheets in Collaborative Design of Knowledge Graphs // Joint Proceedings of the BIR 2018 Short Papers, Workshops and Doctoral Consortium co-located with 17th International Conference Perspectives in Business Informatics Research (BIR 2018). Stockholm, Sweden Vol. 2218. - 2018. - P.82-93. http://ceur-ws.org/Vol-2218/paper8.pdf (Scopus)
12. Kudryavtsev D., Gavrilova T., Leshcheva I., Begler A., Kubelskiy M., Tushkanova O. A method for collaborative visual creation of a knowledge graph // Conference Proceedings "GSOM Emerging Markets Conference 2018" / St.Petersburg, Russia (October 4-6, 2018). - P. 98-102. (РИНЦ)
13. Leshcheva I., Leshchev D. Ontology as mapping of material world // in A. Sarygulov, V. Sergeev, L. Ungvári and W. Semmler (Eds.) SHS Web of Conf. 44 00055 (2018) Proceedings of IV International Scientific Conference “The Convergence of Digital and Physical Worlds: Technological, Economic and Social Challenges” (CC-TESC2018), St. Petersburg, Russia, May 16-18, 2018. - P. 1-9. URL: https://www.shs-conferences.org/articles/shsconf/pdf/2018/05/shsconf_cc-tesc2018_00055.pdf DOI: 10.1051/shsconf/20184400055 (Scopus)

РИНЦ - 9, Scopus – 6, WoS - 2

Список использованной литературы

1. Балацкий Е. В., Юревич М. А. Моделирование возрастной структуры научных кадров // Terra Economicus. – 2018. – Том 16. – №3. – С. 60-76.
2. Бова В. В. Онтологическая модель интеграции данных и знаний в интеллектуальных информационных системах //Известия Южного федерального университета. Технические науки. – 2015. – №. 4 (165). – С. 225-237.
3. Бодрова Е. В., Голованова Н. Б. Модернизация высшей технической школы: исторический опыт и современность // Российский технологический журнал. – 2017. Т. 5. – № 6 (20). – С. 73-97.
4. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям. СПб: Питер, 2013. – 704 с.
5. Aussenac-Gilles N., Kamel M. Ontology Learning by Analyzing XML Document Structure and Content //KEOD. – 2009. – С. 159-165.
6. Bakkas J., Jakjoud W., Bahaj M. Semantic mapping at the schema level of XML documents to ontologies //Next Generation Networks and Services (NGNS), 2014 Fifth International Conference on. – IEEE, 2014. – С. 165-169.
7. Chungoora, N.; Cutting-Decelle, A.-F.; Young, R.I.M.; Gunendran, G.; Usman, Z.; Harding, J.A.; Case, K. Towards the ontology-based consolidation of production-centric standards // International Journal of Production Research. – 2013. – Vol. 51. – Issue 2. – pp. 327-345.
8. Compton M. et al. Declarative Data Exchange: Spreadsheets to RDF/OWL //Proceedings of the 19th International Database Engineering & Applications Symposium. – ACM, 2015. – С. 96-105.

9. Erdmann M., Studer R. How to structure and access XML documents with ontologies //Data & Knowledge Engineering. – 2001. – T. 36. – №. 3. – C. 317-335.
10. Hacherouf M., Bahloul S. N., Cruz C. Transforming XML documents to OWL ontologies: A survey //Journal of Information Science. – 2015. – T. 41. – №. 2. – C. 242-259.
11. Hevner A., Chatterjee S. Design research in information systems: theory and practice. – Springer Science & Business Media, 2010. – T. 22.
12. Michel F., Montagnat J., Faron-Zucker C. A survey of RDB to RDF translation approaches and tools: Research Report. – 2014. URL: http://hal.archives-ouvertes.fr/hal-00903568
13. Nederstigt L. J. et al. Floppies: A framework for large-scale ontology population of product information from tabular data in e-commerce stores //Decision Support Systems. – 2014. – Vol. 59. – pp. 296-311.
14. O'Connor M. J., Halaschek-Wiener C., Musen M. A. Mapping master: a flexible approach for mapping spreadsheets to OWL //International Semantic Web Conference. – Springer, Berlin, Heidelberg, 2010. – C. 194-208.
15. Österle H. et al. Memorandum on design-oriented information systems research //European Journal of Information Systems. – 2011. – T. 20. – №. 1. – C. 7-10.
16. Pan J. Z. et al. (ed.). Exploiting Linked Data and Knowledge Graphs in Large Organisations. – Cham : Springer, 2017. 266 p. doi: 10.1007/978-3-319-45654-6
17. Paulheim H. Knowledge graph refinement: A survey of approaches and evaluation methods //Semantic web. – 2017. – Vol. 8. – №. 3. – pp. 489-508.
18. Peffers K. et al. A design science research methodology for information systems research //Journal of management information systems. – 2007. – T. 24. – №. 3. – C. 45-77.
19. Petasis G. et al. Ontology population and enrichment: State of the art // Knowledge-driven multimedia information extraction and ontology evolution. – Springer-Verlag, 2011. – pp. 134-166.
20. Quix C., Kensche D., Li X. Matching of ontologies with XML schemas using a generic metamodel //OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". – Springer, Berlin, Heidelberg, 2007. – C. 1081-1098.
21. Sitharamulu V., Babu B. R. A Novel Proposal for Bridging Gap between RDB-RDF Semantic Web using Bidirectional Approach //International Journal of Applied Engineering Research. – 2016. – Vol. 11. – №. 6. – pp. 4456-4460.
22. Thuy P. T. T. et al. Exploiting XML schema for interpreting XML documents as RDF //Services Computing, 2008. SCC'08. IEEE International Conference on. – IEEE, 2008. – T. 2. – C. 555-558.
23. Tijerino Y. A. et al. Towards ontology generation from tables //World Wide Web. – 2005. – T. 8. – №. 3. – C. 261-285.
24. Uschold M. Demystifying Owl for the Enterprise (Synthesis Lectures on Semantic Web: Theory and Technology) //Morgan & Claypool Publishers, 2018. – 237 p.
25. Valarakos A. G. et al. Enhancing ontological knowledge through ontology population and enrichment //International Conference on Knowledge Engineering and Knowledge Management. – Springer, Berlin, Heidelberg, 2004. – pp. 144-156.

Участие в научных мероприятиях по тематике проекта за 2018 год

1. 21ая Российская научная конференция «Инжиниринг предприятий и управление знаниями» (25-28 апреля 2018 г.), ФГБОУ ВО «РЭУ им. Г.В. Плеханова» / г. Москва, Россия.
2. 4th International Scientific Conference “The Convergence of Digital and Physical Worlds: Technological, Economic and Social Challenges” (CC-TESC2018) (16-18 May, 2018) Peter the Great St. Petersburg Polytechnic University / St. Petersburg, Russia (4ая Международная научная конференция «Конвергенция цифровых и физических миров: технологические, экономические и социальные проблемы» (CC-TESC2018) (16-18 мая 2018г.), Санкт-Петербургский политехнический университет Петра Великого / г. Санкт-Петербург, Россия).
3. 13th International Forum on Knowledge Asset Dynamics «IFKAD 2018» (4-6 July, 2018), Delft University of Technology / Delft, Netherlands (13ый Международный Форум динамики знаниевых активов «IFKAD 2018» (4-6 июля 2018 г.), Делфтский технический университет / г. Делфт, Нидерланды).
4. 19th European Conference on Knowledge management «ECKM» (6-7 September, 2018), University of Padua / Padua, Italy (19ая Европейская Конференция по Управлению Знаниями «ECKM» (6-7 сентября 2018 г.), Падуанский университет / г. Падуя, Италия).
5. 17th International Conference on Perspectives in Business Informatics Research (BIR 2018 – CEUR Workshop Proceedings) (24-26 September, 2018), Stockholm University / Stockholm, Sweden (17ая Международная конференция по перспективам исследований бизнес-информатики (BIR 2018 – Семинар-практикум CEUR) (24-26 сентября 2018 г.), Стокгольмский университет / г. Стокгольм, Швеция).
6. 16ая Национальная конференция по искусственному интеллекту с международным участием (КИИ-2018) (24 - 27 сентября 2018 г.), Федеральный исследовательский центр «Информатика и управление» РАН (ФИЦ ИУ РАН) совместно с Российской ассоциацией искусственного интеллекта (РАИИ), Высшей Школой Экономики (НИУ ВШЭ), Институтом проблем управления РАН (ИПУ РАН) и The Institute of Information and Communication Technologies at the Bulgarian Academy of Sciences / г. Москва, Россия.
7. 5th International Research GSOM Emerging Markets Conference 2018 (4-6 October, 2018), Graduate School of Management SPbU / St.Petersburg, Russia (5ая Научная Международная Конференция «Развивающиеся рынки 2018» (4-6 октября 2018 г.), Высшая Школа Менеджмента СПбГУ / г. Санкт-Петербург, Россия).
8. 5ая Международная научно-практическая конференция “Управленческие науки в современном мире” 2018 (7-9 ноября 2018 г.), Финансовый Университет при Правительстве РФ / г. Москва, Россия.