2013 | Volume 16| Issue 6|

Интеграция данных по свойствам веществ и материалов на основе онтологического моделирования предметной области

А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, В.А. Серебряков, Л.Н. Шиолашвили

Аннотация

Изучена специфика предметной области "Свойства веществ и материалов", рассматриваемой как "полигон" при отработке подходов к интеграции научных Баз данных. На конкретных примерах показаны возможности онтологического моделирования для интеграции численных данных по свойствам на семантическом уровне. Предложена общая схема онтологии для представления термодинамических свойств данных по свойствам чистых соединений.

Ключевые слова: онтологическое моделирование, формализация предметной области, интеграция численных данных, термодинамические свойства вещества.

1. Введение

Опыт множества прикладных проектов показал, что при разработке информационных технологий наиболее практичным оказывается подход ***bottom-up***, сужающий исходные понятия и структуры для ограниченной предметной области при сохранении потенциала расширения тематики и охвата все новых ресурсов. Поэтому, выделяя для отработки технологий некоторую область естественных или инженерных наук (скажем, молекулярная физика, кристаллография, химия природных соединений и проч.), оправдано сузить тематику, концентрируя внимание только на данных по свойствам. С одной стороны, эти данные всегда одним из продуктов научной деятельности, но с другой их структура и организация заведомо проще, чем вся совокупность научного знания, включающая анализ процессов, открытие закономерностей, обширную сферу приложений и т.п. Представляется, что именно тематику **«свойства вещества…»»** можно предложить как **полигон** при отработке концепций и технологий интеграции научных баз данных (БД). Создав на ограниченном плацдарме адекватную модель интеграции данных, можно надеяться на дальнейшее расширение ее возможностей, с целью охвата информационных ресурсов произвольной направленности.

Есть несколько доводов в пользу выбора тематики **«свойствами веществ…»** при развитии методов интеграции:

- работы этого направления представляют интерес для всего научного

сообщества, включенного в решение проблем физики, химии, материаловедения, а частично биологии и наук о Земле;
- научная активность в этих областях неразрывно связана с созданием и поддержкой многообразных БД, что порождает высокую заинтересованность в развитии новых технологий, облегчающих решение проблем стандартизации и обмена научными данными;
- при крайней широте тематики по своему генезису данные, относящихся к тематике «свойства вещества», естественным образом соответствуют типовым структурам реляционных БД. В каждой из конкретных областей всегда выделяют перечень объектов (веществ, молекул и т.п.), задают перечень свойств и набор метаданных, раскрывающих структуру, форматы и, частично, семантику данных;
- налицо уже сейчас определенные результаты, свидетельствующие о возможности интеграции данных в достаточно широкой предметной области веществ и материалов при правильно поставленной формализации понятий и связей.

2. Структура данных

Во множестве дисциплин появились собственные версии XML со своими словарями, средствами поддержки в виде настраиваемых браузеров и программ, реализующих графические представления, вычислительные сервисы и проч. [1] К настоящему времени созданы десятки таких версий, списки которых можно найти на сайтах www.xml.com/pub/rg/Science или xml.coverpages.org/xmlApplications.html. По крайней мере, два из них (ThermoML и MatML) оказались достаточно успешны при распространении и обмене данными по свойствам в термодинамике и материаловедении. В определенной степени этот успех связан с тем, что ключевые данные в обеих областях имеют относительно простую структуру: объекту с характерным для него именем (или набором имен) приписывается некоторый набор свойств в виде констант или одномерных таблиц.

Проект ThermoML – разработка Термодинамического Центра Института стандартов и технологий США [2], в качестве основной цели имеет стандартизацию форм представления и обмена теплофизическими данными. Язык передает данные для более чем 120 свойств при различных формах представления и различном статусе данных (экспериментальные, расчетные, справочные). В качестве объектов могут рассматриваться чистые вещества с характерной для них идентификацией (по формуле, названию, номеру в принятой номенклатуре), смесь, определяемая данными по составу, химические и фазовые реакции. Для представления данных разработана детальная схема, доступная в сети, http://www.trc.nist.gov/ThermoML.xsd.

Структура ThermoML-документа представляет собой сбалансированную комбинацию иерархических и реляционных элементов. В явной форме схема вводит все понятия, названия свойств и их классов, параметров состояния, ограничений, фаз. Большое место занимают метаданные, детализирующие метод измерения, состояние образца, форму представления неопределенности. Предусмотрено использование MathML для передачи данных в виде формул. Проект обеспечен специальными средствами, позволяющими выделять из научных

статей данные по свойствам и генерировать соответствующие XML-документы. На сайте NIST (http://trc.nist.gov/ThermoML.html) приведены примеры трансформации статей из группы физико-химических журналов в соответствующие XML-файлы. Детальный анализ возможностей ThermoML при интеграции разнообразных теплофизических данных приведен в работе авторов [3].

3. Онтология как средство представления справочных данных

3.1. Онтологическое моделирование в науке о материалах

Хотя язык XML и его профессиональные версии уже сейчас способны обеспечить интероперабельность в среде гетерогенных БД и приложений, они не представляют семантику (смысл) распространяемых данных. Для этих целей требуется более высокий уровень интеграции, «задуманный» в известной статье Тима Бернерса-Ли и др. [4], предложивших концепцию ***Semantic WEB***. Строение Semantic Web напоминает по выражению Бернерса-Ли «слоеный пирог», включая на нижнем уровне XML для определения схемы данных, затем RDF (Resource Definition Framework) для метаданных, и наконец, на верхнем уровне онтологии.

Подлинным ядром Semantic Web является **онтология** - система понятий предметной области, которая представлена как набор сущностей, соединенных различными отношениями. Именно онтология передает знания в виде формальной структуры, доступной для компьютерной обработки. В 2004 году World Wide Web Consortium (W3C) предложил универсальный стандарт для сетевого обмена онтологической информацией – Web Ontology Language (OWL).

Применительно к данным о свойствах вещества есть несколько удачных примеров использования онтологического моделирования. В основном, они относятся к области материаловедения, где многообразие типов данных и богатство словарей проявляются наиболее ярко. Среди таких примеров база знаний PLINIUS, оперирующая результатами исследований по свойствам керамик [5], онтологическое описание свойств ползучести конструкционных материалов [6], система MatONT [7], спроектированная для поддержки исследований по новым материалам. Примерно той же цели, но с охватом более широкой информации, включающей, наряду с материалами, промышленные изделия, служит стандарт ISO 10303-235: **«Engineering properties for product design and verification»** [8]. Стандарт предусматривает единую информационную модель для определения семантики и синтаксиса представления и единый словарь для определения смысла данных.

В наиболее общем виде методика онтологического описания данных по свойствам сформулирована в работе [9]. Автор обращается к концепции **Semantic Web**, используя слоистую структуру со стандартизованными процедурами перехода от нижнего слоя к верхнему. Нижний слой включает XML для определения схемы данных, средний слой (RDF) для определения метаданных и наконец, OWL для представления онтологий, таблица 1. Существенно, что по отдельности эти элементы давно разработаны и стандартизованы. В сравнении с MatML, подобная

структура способна обеспечить более высокий уровень стандартизации, формализующей определение свойств, методов обработки и использования.

Таблица 1. Структура Semantic Web для данных по свойствам материалов.

Слой	**Содержание**
Онтологии (OWL)	Таксономия материалов и свойств
Метаданные (RDF)	Метаданные, использованные в БД
Схема (XML Schema)	Схема данных по свойствам материалов (MatML)

Онтология, покрывающая предметную область «свойства материалов», согласно [9], должна включать 7 онтологий, распределенных по 3 группам - три столбца в таблице 2.

Таблица 2. Общий состав онтологии для описания материалов

Базовые онтологии	**Информация по материалу**	**Вспомогательные онтологии**
Вещество		Единицы измерения
Процесс		Физические константы
Свойство		
Окружение		

Четыре базовые онтологии дают определения терминов, названий и словарей, представляющих основные концепции для каждой из областей. Каждая из онтологий основана на таксономии классов, представленных в словаре понятий. Пример таксономии классов для свойств и самих материалов дают рисунки из статьи, приведенные ниже.

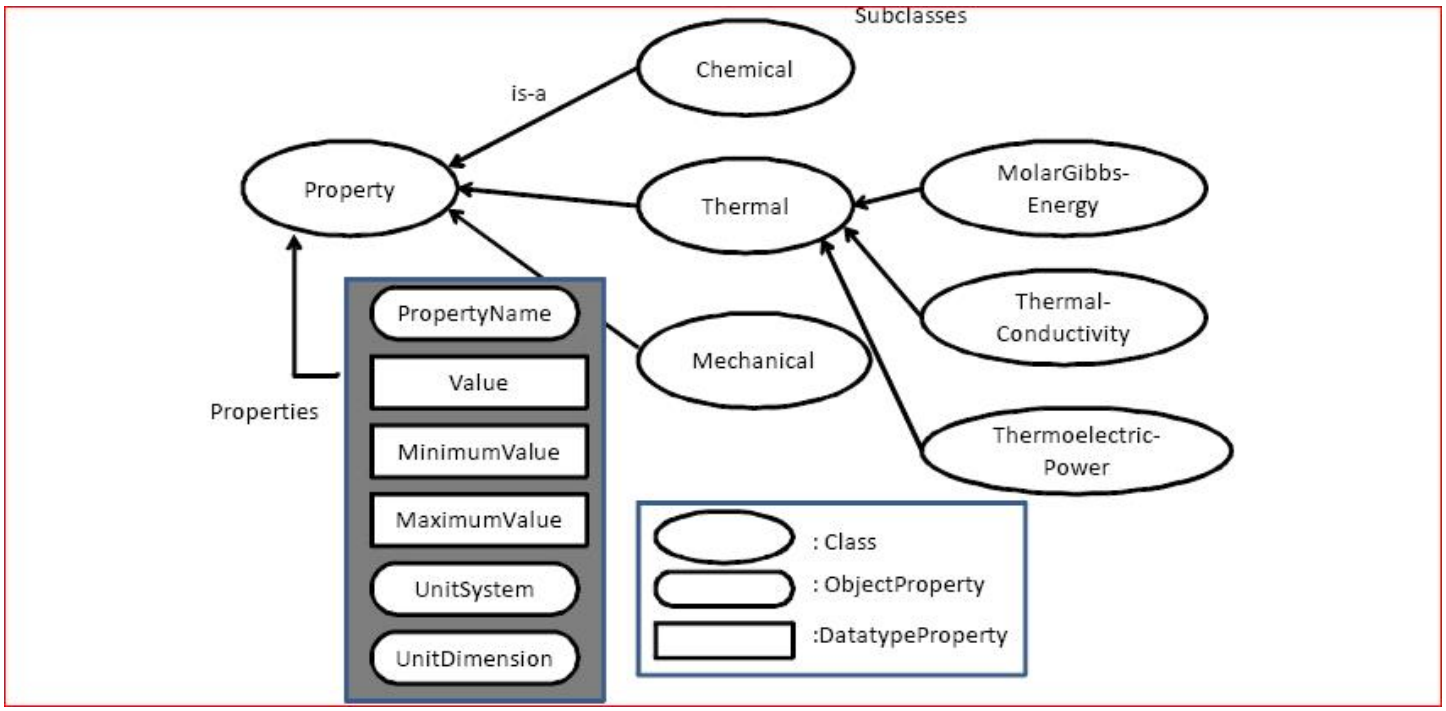

Рис. 1. Структура класса Property в онтологии Ashino [9]

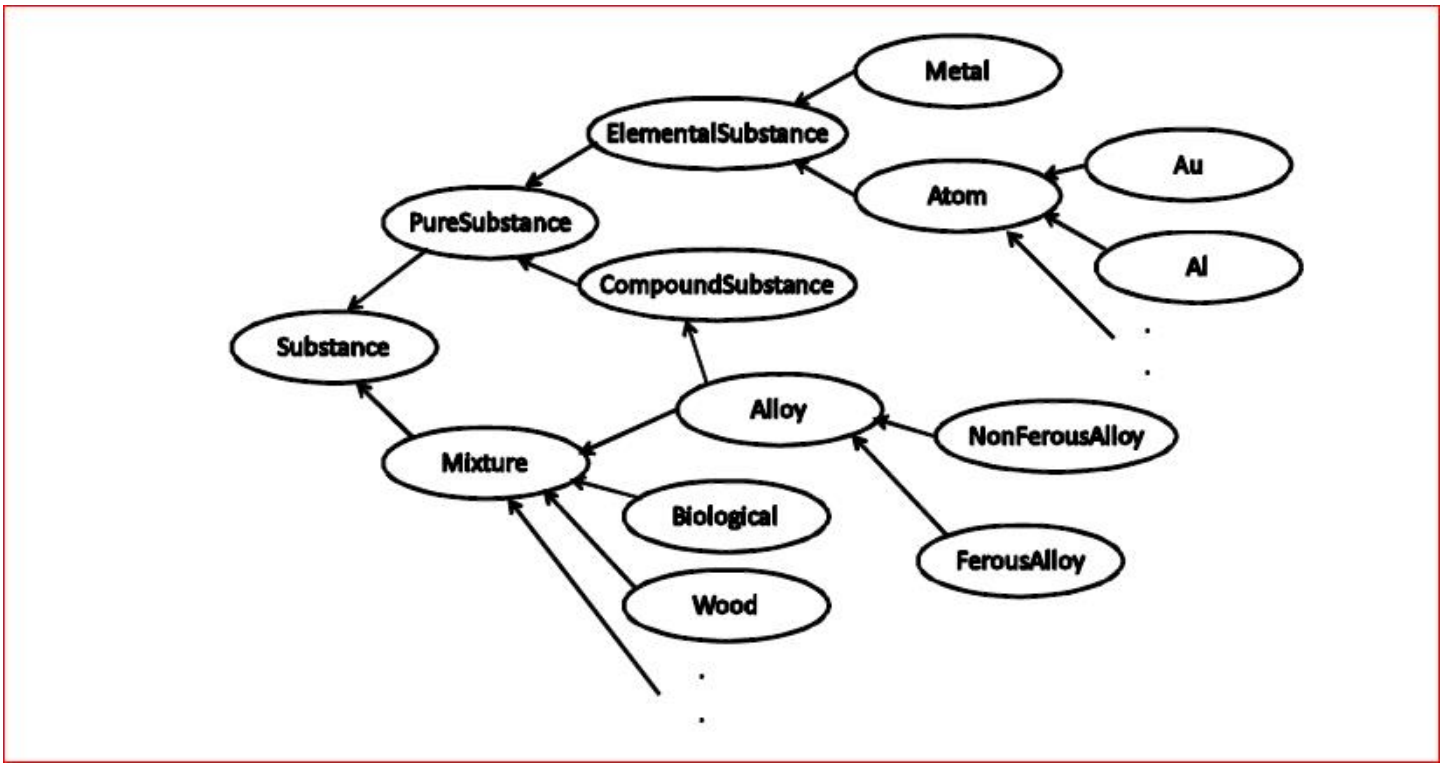

Рис. 2. Таксономия веществ в классе Substance [9]

Две другие базовые онтологии («процесс» и «окружение»), отмеченные в таблице 2, дают описания соответственно методов производства и измерения и характеристик среды: состав атмосферы, температуру, pH и т.п. Сверх четырех базовых, в общую онтологию включена «информация по материалу», детализирующая сведения по конкретному объекту путем агрегирования других классов (вещество, свойство и т.п.).

Используя базовые онтологии, эта онтология агрегирует все термины и концепции, характеризующие материал и конкретный образец, методы и условия измерения, критерии качества данных и проч. Согласно таблице 1 онтология по материалам включает и вспомогательные онтологии, определяющие общенаучные концепции. В частности, при построении онтологии «единицы измерений» используется синтаксис MathML (версия, предназначенная для передачи формул), чтобы ввести операции, необходимые при согласовании различных единиц измерения.

Онтология по материалам прошла тестирование на типовой процедуре обмена данными – среди группы разнородных БД, содержащих информацию по теплофизическим свойствам. Процедура сводится к конверсии логической структуры каждой из БД в единую структуру, предусмотренную разработанной онтологией. Таким образом, каждая из реляционных БД по свойствам может экспортировать данные в едином формате, пригодном как для обмена, так и для долгосрочной архивации. Отображение полей каждой их БД производится посредством XSLT шаблонов, причем первичную подстройку в полной степени автоматизировать не удается.

Предложенная в [9] онтология представляет собой структурированный словарь понятий и концепций, принятых в науке о материалах, встроенный в рамки Semantic Web. Благодаря использованию OWL, выделение общих понятий обеспечено, в основном, пространством имен (namespaces) и идентификаторами URI. Онтология по материалам предоставляет общие для всех ресурсов термины и нотации для манипуляции данными и знаниями. Еще один компонент онтологии – цифровая библиотека эмпирических/теоретических уравнений, записанных на MathML, языке математической разметки. Все компоненты онтологии используют общий формат данных (XML) и могут быть размещены в сети интернет.

3.2. Интеграция с онтологиями общенаучного содержания

Проекты онтологий, развитые в работе [9], как и некоторых других [5-8] в основном фокусируются на деталях конкретных дисциплин, не предполагая возможность расширяемости. Значительно более широкие возможности интеграции предусмотрены концепцией проекта MatOnto. Онтология под этим названием предоставляет экспертам и разработчикам БД платформу для интеграции существующих и возникающих дисциплин в рамках науки о материалах. При ее построении использовано довольно большое количество уже существующих онтологий и классификационных схем: Ontolingua's Standard Units and Dimensions; Joint Academic Classification of Subjects (JACS); W3C’s Time Ontology; AIFB’s Semantic Web для исследовательского сообщества. Обеспечена также связь с онтологией EXPO для описания научных экспериментов вместе с онтологией метаданных ABC, расширяющей EXPO концепциями событий и процессов.

На рис. 3 и 4 из работы авторов проекта MatOnto схематично представлен верхний уровень онтологии. На первом из них показана связь с другими онтологиями, на рис. 4 – возникает ключевая онтологии – *matonto:Material*. С ней связаны 5 базовых онтологий, определяющих все аспекты работы с материалами.

matonto:Property	**Свойства материалов**
matonto:Family	Классификация материалов
matonto:Process	Процессы роизводства и тестирования
matonto:Structure	Структура материалов
matonto:Measurement	Данные измерений или процесса идентификации материала

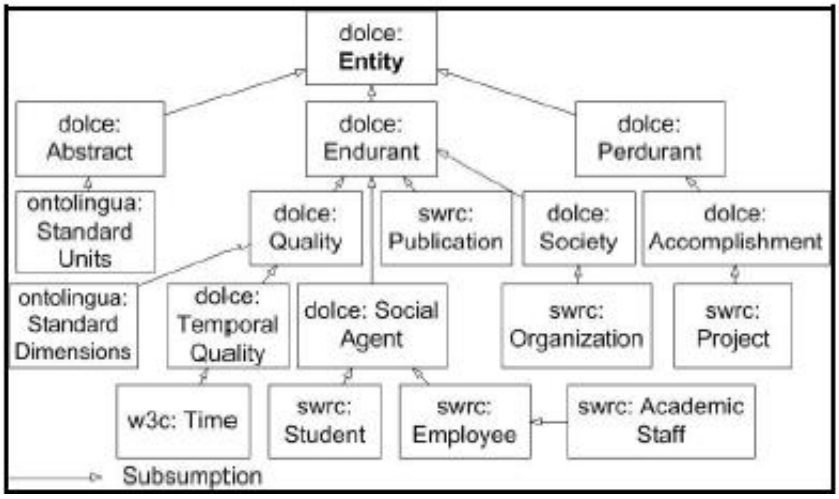

Рис. 3. Классы верхнего уровня в онтологии MatOnto

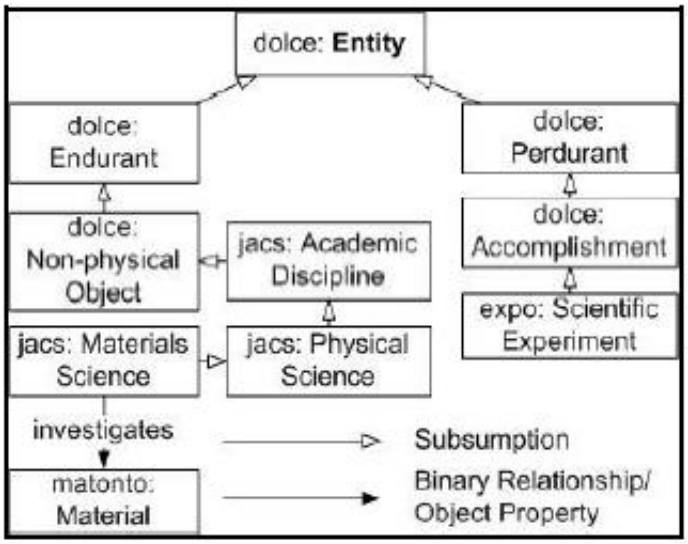

Рис. 4. Адаптация классов верхнего уровня к задачам науки о материалах

4. Проект онтологии для представления данных по термодинамическим свойствам

Даже поверхностный анализ работ [5-7, 9] показывает, что разработчики онтологий по материалам ставят перед собой слишком сложные задачи, по сути, формализацию всей дисциплины, включая самые разные аспекты: постановка экспериментов, производственный процесс, научная активность (например,

выдвижение гипотез, планирование и обработка экспериментов и т.п.) – см. схемы на рис. 5 и 6 из статьи [7].

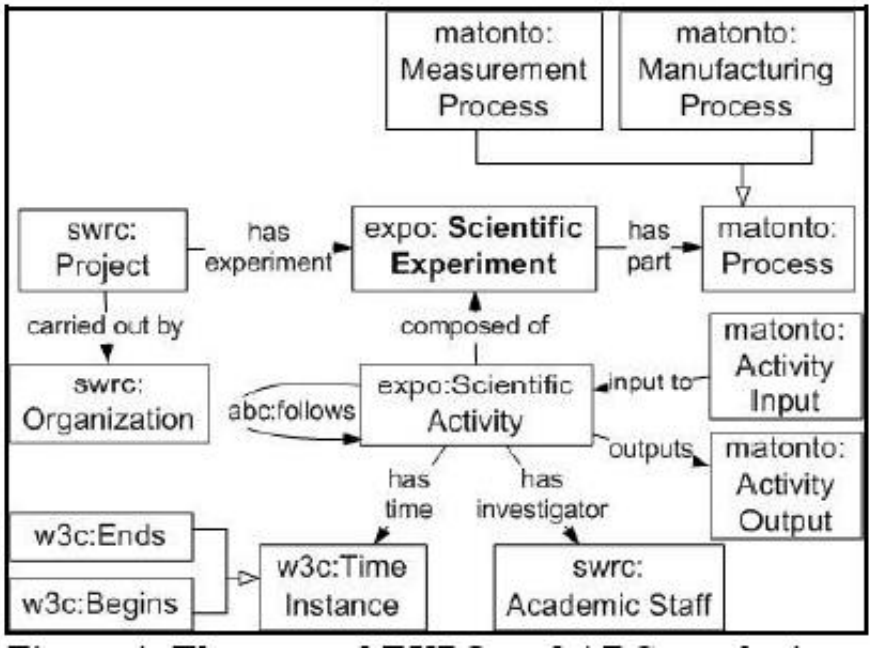

Рис. 5. Использование peer-reviewed онтологии EXPO для описания научных экспериментов (в сочетании с ABC Metadata Ontology , позволяющей вволить концепции событий и процессов).

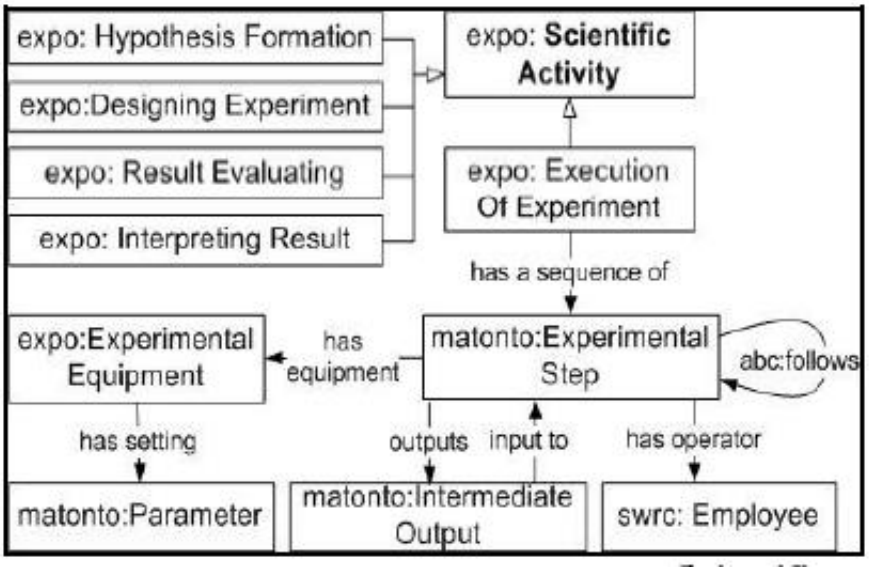

Рис. 6. Класс Scientific Activity для детального описания постановки эксперимента

На наш взгляд, подобная формализация излишне усложняет реальную задачу, которая ограничена стандартизацией данных по свойствам объектов с возможностью их интеграции в среде Баз данных с различной структурой и семантикой.

Более того, на этапе отработки технологий оправдано ограничить круг объектов, например, «веществами», а не «материалами». Принципиально различие между

этими терминами в том, что под «материалом» понимается объект, свойства которого в заметной степени определяются сферой производства: завод-изготовитель, марка, технология, условия хранения и поставки и проч. Все это увеличивает объем сведений и резко усложняет формализацию концепций.

В качестве альтернативы мы предлагаем ограничиться на начальном этапе именно «веществами», свойства которых определяются, в основном, их природой: стехиометрической формулой, составом, фазой и проч. Небольшой объем деталей, уточняющих данные для конкретного образца (сведения о его чистоте, внешних воздействиях и т.п.) принято в модели данных инкапсулировать в блоке «SAMPLE» (образец), само наличие которого необходимо только при распространении экспериментальных данных [10]. Более того, при разработке онтологии для передачи данных по свойствам, мы сочли возможным ограничиться пока чистыми веществам (исключая смеси, растворы, сплавы) и теми свойствами, данные по которым можно представить либо константами, либо функцией одной-двух переменных, чаще всего, в виде одномерных таблиц. Рассмотрим примерную схему, определяющую верхний уровень требуемой онтологии, показанную в виде UML-диаграммы на приведенном ниже рис. 7.

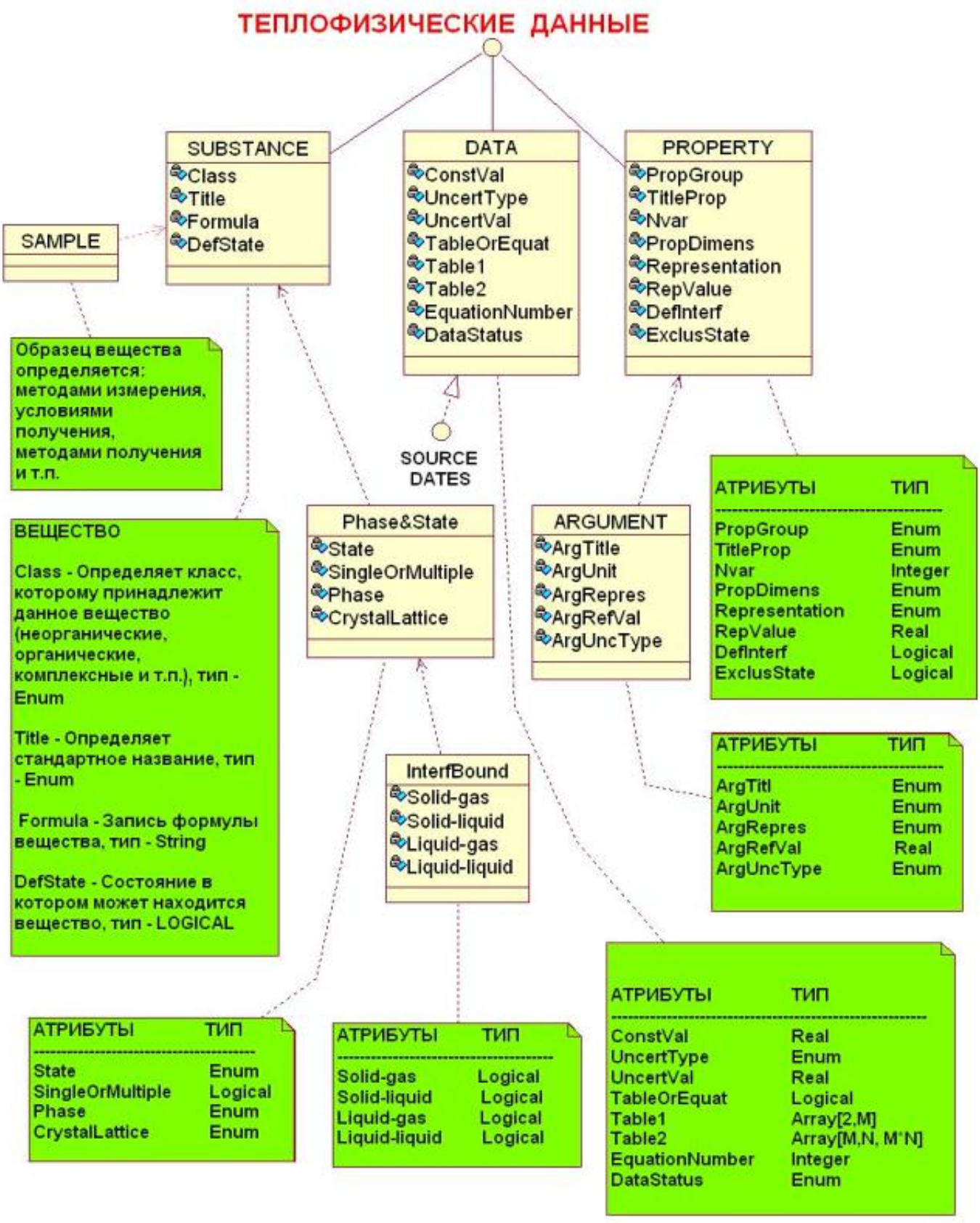

Рис. 7. UML-диаграмма онтологии

Таблица 3. Классы онтологии для представления теплофизических свойств.

Основные классы онтологии	**Вспомогательные классы**

SUBSTANCE	SAMPLE
PROPERTY	SOURCE
ARGUMENT	
DATA	
Phase&State	
InterfBound	

На диаграмме вспомогательные понятия не раскрыты, а для каждого из основных указаны их атрибуты (свойства) и представляющий их тип данных. Большинство атрибутов передаются посредством классификаторов (тип данных **Enum**). Например, первый из классификаторов, раскрывающий понятие **SUBSTANCE**, выделяет класс веществ: органические, неорганические, комплексные и т.п. Следующий классификатор **title** позволяет в любом из указанных классов выделить по названию определенное вещество. Наряду **Enum**, используются также другие типы данных: **string, real, integer, logical**.

Три главных класса в онтологии **(SUBSTANCE, PROPERTY, DATA)** определяют вещество, свойство и сам набор данных **(DATA)**. Первый выделяет конкретное вещество, второй – определенное свойство из перечня, предполагая детализацию свойств с выделением термических, калорических, транспортных, механических и др. Разумеется, класс свойствам должен быть связан со словарем, включающим единицы измерений. Экземпляр класса детализирует форму представления и включает сами численные данные (плотность, теплоемкость и проч.), а также сведения о неопределенности.

Три указанных класса было бы достаточно для простейшей модели данных, когда каждому веществу можно приписать каждое из возможных свойств, а для любой пары «вещество-свойство» задать набор данных. Реальная картина, связанная как с таксономией характеристик, так и физическими ограничениями, требует ввести в рассмотрение ряд других классов. Прежде всего, вещество может пребывать в различных состояниях, что отражается на используемом перечне понятий. Например, выделяя три простейших состояния (твердое, жидкое и газообразное), мы должны помнить, как это отражается на номенклатуре свойств. Скажем, понятие «вязкость» неприменимо к твердому состоянию вещества, а понятие «модуль Юнга» неприменимо к жидкому и газообразному. С другой стороны, и среди веществ можно выделить подгруппы, с ограниченным отнесением к разным состояниям и фазам. Так, если сравнить несколько веществ, таких как , только первые три – обычные вещества, которые можно рассматривать в разных фазовых состояниях, приписывая им все свойства, которые им соответствуют. Последний из списка, гидроксил , радикал, который может рассматриваться только как компонент идеально-газовой смеси. Применительно к нему теряет смысл представление о разных фазах и состояниях или таких типичных характеристиках вещества как температура кипения, уравнение состояния и проч.

Учесть всю сложность, связанную с подобными ограничениями, введено два класса **Phase&State** и **InterfBound**: первый для выделения состояний и фаз, в

которых может находиться вещество, второй – для выделения межфазных границ. Распределяет вещество между этими классами атрибут **SingleOrMultiple**, который при значении **«Yes»** указывает, что достаточно указать одно состояние и фазу, а при значении **«No»** переводит рассмотрение в класс **Interf&Bound**, где выделяется одна из фазовых границ (например, линия насыщения – жидкость-газ).

Некоторые связи и ограничения, возникающие при образовании связки «вещество-свойство», обеспечивают атрибуты в классах **SUBSTANCE** и **PROPERTY**. К примеру, **DefState** при значении **«Yes»** указывает на то, что вещество может находиться только в одном состоянии, как правило, состоянии идеального газа.

Аналогично атрибуты **DefInterf** и **ExclusState** накладывают ограничения на выбор вещества, к которому может относиться свойство. Первый при значении **«Yes»** указывает, что надо выделить одну из фазовых границ в классе **Interf&Bound**, чтобы присвоение свойства веществу было законным. Второму присваивается значение **«Yes»** , если надо исключить возможность присвоения данного свойства данному веществу.

Конкретный набор численных данных дает экземпляр класса **DATA.** Форма представления зависит от значения Nvar в классе PROPERTY. Если Nvar=0, задается значение константы; если Nvar=1 или 2 задается функция в виде таблицы или формулы. В двух последних случаях свойство рассматривается как функция одного или двух аргументов, скажем температуры и давления. Требуемые характеристики аргумента/ов определяют экземпляры класса **ARGUMENT**.

Два атрибута **UncertType** и **UncertVal** задают форму представления неопределенности (абсолютная, относительная и т.п.) и само значение неопределенности. Предусмотрено в классе **DATA** также разделение данных на 3 категории: экспериментальные, расчетно-теоретические, справочные. Только для первых учитываются сведения, предусмотренные классом **SAMPLE**.

Несмотря на сужение предметной области, онтология оказывается достаточно сложной за счет множества ограничений, накладываемых на связи между понятиями. Есть ограничения между веществом, свойством и теми состояниями и фазами, которые определяют состояние вещества. Примеры таких ограничений представлены на рис. 8 и 9.

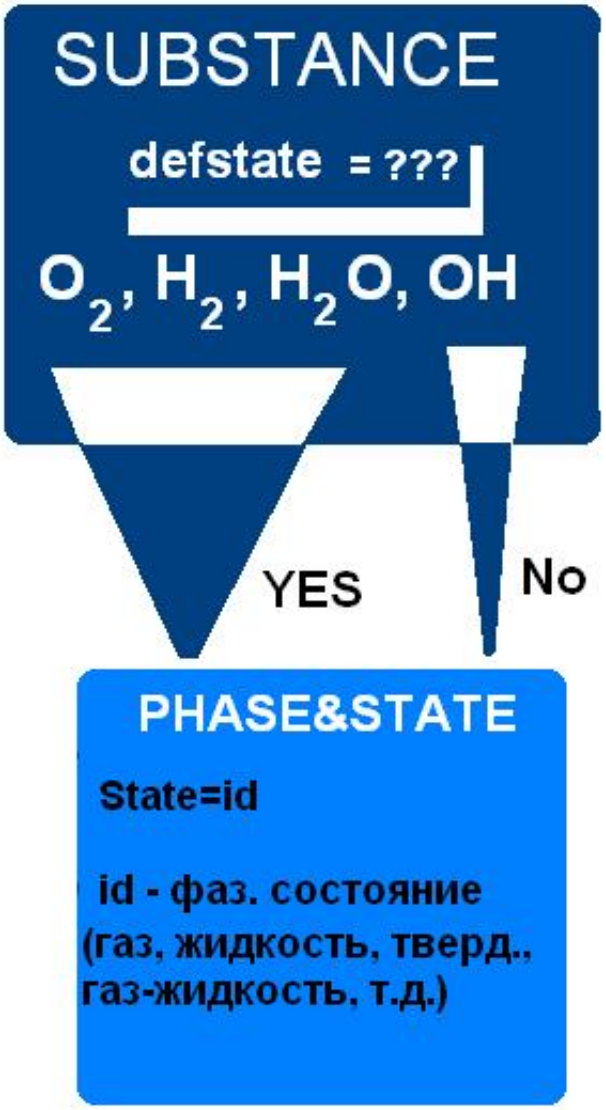

Рис. 8. Ограничивающая связь между выбором вещества и состояния.

Рис. 9. Ограничивающая связь между выбором свойства и отнесением состояния к

пограничной линии.

Другие ограничения связаны с параметром Nvar. Если Nvar =0 нет обращений к классу **ARGUMENT**; если Nvar =1 заполняются данные для одного экземпляра класса **ARGUMENT**, если Nvar =2 заполняются данные для двух экземпляров класса **ARGUMENT**. В двух последних случаях слот **TableOrEquat** определяет заполнение таблицы или обращение к формуле, представляющей зависимость свойства от параметра состояния.

Общее число подобных ограничений, связанных с адекватным отражением предметной области достаточно велико. Помимо необходимости учитывать подобные логические ограничения (типа «если….то»), приходится учитывать и те, которые связаны с физическими принципами. В качестве примера, можно привести связь трех функций и аргумента, приводимых в термодинамических справочниках

$F = S -$

$\underline{H}$

T

где F — энергия Гиббса, H — энтальпия, S — энтропия, T — параметр состояния (температура).

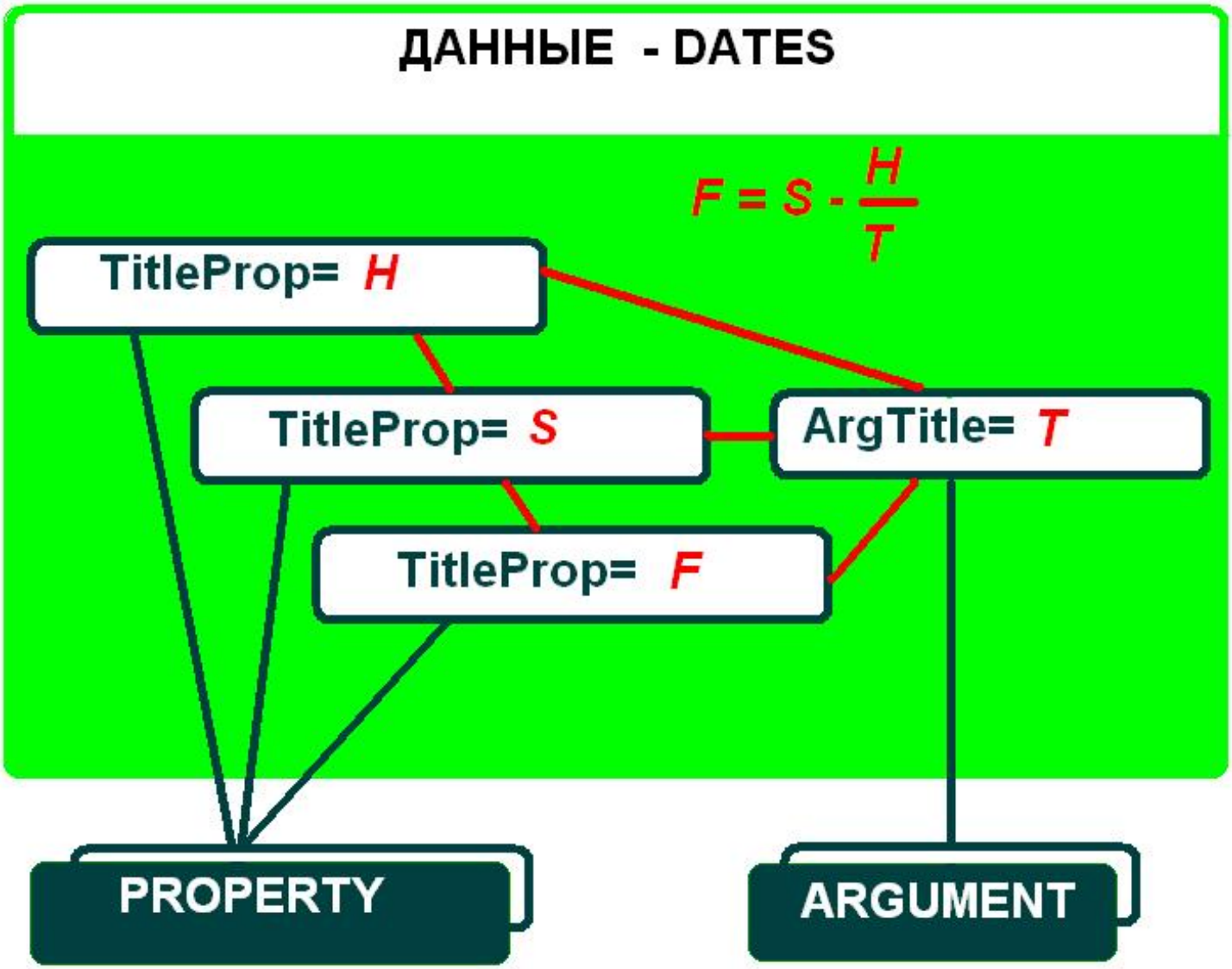

Рис. 10. Ограничивающая связь между разными экземплярами класса PROPERTY.

Задача данного проекта – довести формализацию предметной области до уровня, позволяющего для «суженной» предметной области охватить практически все виды представляемых в литературе данных по свойствам. Окончательная задача – разработка софта, обеспечивающего экспорт данных из БД или публикаций в форме, соответствующей разработанной онтологии.

Литература

1. А.О Еркимбаев., В.Ю. Зицерман, Г.А. Кобзев. Версии языка XML в задачах хранения и распространения научных данных. Сборник трудов Всероссийской научной школы-семинара молодых ученых, аспирантов и студентов, "Интеллектуализация информационного поиска, скантехнологии и электронные библиотеки". Таганрог: Изд-во ТТИ ЮФУ, 2011, стр. 52-58.

2. M. Frenkel. Global communications and expert systems in thermodynamics: Connecting property measurement and chemical process design// Pure Applied Chem. 2005. V. 77. No. 8. PP. 1349 - 1367.

3. А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, Л.Р. Фокин. Логическая структура физико-химических данных. Проблемы стандартизации и обмена численными данными// Журнал физической химии. 2008. Т.82. №1. С. 20-31.

4. T. Berners-Lee, J. Hendler, & O. Lassila. The Semantic Web// Scientific American 2001. V. 284(5). PP. 35-43.

5. Van der Vet, P. E., Speel, P-H., & Mars, N. J. I. Ontologies for very large knowledge bases in materials science: A case study// The Second International Conference on Building and Sharing Very Large-Scale Knowledge Bases, 1995, University of Twente, 73-83.

6. T. Ashino, M. Fujita. Definition of Web Ontology for Design-Oriented Material Selection. Data Science Journal 2006, V. 5, pp. 52-63.

7. Cheung, K., Drennan, J., & Hunter , J. (2008) Towards an Ontology for Data-driven Discovery of New Materials. AAAI Workshop on Semantic Scientific Knowledge Integration, Stanford University, 26-28.

8. N. Swindells. The representation and exchange of materials and other engineering properties. Data Science Journal. 2009. V. 8, PP. 190-200.

9. T. Ashino. Materials ontology: an infrastructure for exchange materials information and knowledge. Data Science Journal, Volume 9, 8 July 2010, pp. 54-61.

10. A. Kazakov, C.D. Muzny, K. Kroenlein, V. Diky, R.D. Chirico, J.W. Magee, I.M. Abdulagatov, M. Frenkel. NIST/TRC SOURCE Archival System: The Next Generation Data Model for Storage of Thermophysical Properties. International J. Thermophys. 2012. V. 33, P. 22-33.

Об авторах

Еркимбаев А.О. – к.т.н., зав. Информационно-вычислительным центром Объединенного Института высоких температур (ОИВТ) РАН., e-mail: K.Kuznetcov@gmail.com

Зицерман В.Ю. – к.ф-м.н., зав.лабораторией ОИВТ РАН. e-mail: serebr@ccas.ru

Кобзев Г.А. – д.ф-м.н., зав. отделом ОИВТ РАН. e-mail: kbt@ccas.ru

Серебряков В.А. - д.ф-м.н., профессор, зав. отделом Вычислительного центра (ВЦ) РАН. e-mail: kbt@ccas.ru

Шиолашвили Л.Н. - - м.н.с. ВЦ РАН. e-mail: kbt@ccas.ru
