Распространение данных по теплофизическим свойствам в пространстве LOD (Linked Open Data) ♣

В.А. Серебряков[1], К.Б. Теймуразов[1], Е.И. Устинова[2], А.О. Еркимбаев[3], Г.А. Кобзев[3], В.Ю. Зицерман[3]

Вычислительный центр РАН[1] , Факультет ВМК МГУ[2], ОИВТ РАН[3]

serebr@ccas.ru, kbt@intring.ru, jane.echo90@gmail.com, adilbek@ihed.ras.ru, gkbz@mail.ru, vz1941@mail.ru

Аннотация

Разработана детальная онтология для представления и связывания в пространстве LOD наборов численных данных по теплофизическим свойствам веществ. Предусмотрено повторное использование терминов внешних онтологий (ChemAxiom, QUDT) и внесение связей, как логического, так и математического характера. Исходные данные, представленные как записи реляционной базы данных, конвертируются к RDF-формату с помощью свободно распространяемого инструмента D2R. Предложена технология связывания авторского набора данных с внешними ресурсами LOD для веществ, свойств и единиц измерения.

1. Введение

В докладе на прошлой конференции IMS-2012 [1] нашим коллективом была представлена общая концепция интеграции теплофизических данных методами Semantic Web. Была изучена специфика предметной области «Свойства веществ и материалов», и на ее основе предложена общая схема онтологии для представления термодинамических данных по свойствам чистых соединений. За прошедшее время проведена работа по детализации онтологии, путем ее распространения на более широкий круг объектов и свойств, и предложены конкретные технологические решения по генерации наборов данных и их подготовке к публикации в пространстве LOD.

Целью данной работы являлась интеграция данных по теплофизическим свойствам, представленных в термодинамических и теплофизических БД ОИВТ РАН путем их публикации в пространстве LOD и связывания с тематически родственными данными, а также словарями и онтологиями, размещенными в сети.

Вообще говоря, в рамках этой технологии задача интеграции решается независимо для отдельно взятого набора данных, так как процесс ее решения не является строго зафиксированным и может различаться для разных исходных данных и предметных областей. Однако, можно выделить основные этапы интеграции: 1) разработка онтологии предметной области или выбор одной из существующих онтологий; 2) конверсия исходных данных в RDF формат, соответствующий онтологии предметной области; 3) предоставление доступа к полученным данным в соответствии требованиями LOD; 4) генерация типизованных связей между ресурсами полученного набора данных и ресурсами внешних наборов в LOD.

Необходимость «ручной» подгонки данных к требованиям LOD обусловлена общими проблемами, в числе которых: разнообразие терминов и разобщенность данных. Разнообразие терминов проявляется, когда для описания одних и тех же понятий (классов) в рамках предметной области создатели разных наборов ориентировались на разные онтологии. Такая ситуация исключает возможность обработки данных программными агентами, в частности, может привести к невозможности однородного доступа к данным из таких наборов. Под разобщенностью данных понимается отсутствие ссылок на сущности, используемые в других наборах данных, что, влечет появление множества дублирующих ресурсов. Указанные проблемы иллюстрируют отсутствие жестких стандартов, которые могли бы регламентировать процесс интеграции набора данных в пространстве LOD. Поэтому первостепенной задачей, предваряющей процесс публикации в LOD, является построение онтологии предметной области.

Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS-2014, Санкт-Петербург, 19 - 20 ноября 2014 г.

Главный объект в рассматриваемых наборах данных - численные значения свойств для различных веществ при разных условиях среды. Основные сущности, используемые в данной предметной области: вещества, свойства, единицы измерения, наборы численных значений свойств для вещества в определенном состоянии. Вещества характеризуются систематическим названием, химической формулой и могут находиться в различных фазовых состояниях. Фазовые состояния включают в себя набор агрегатных состояний (газ, жидкость, твердое состояние), межфазные границы (газ-жидкость, газ-твердое тело, жидкость-твердое тело) и тройную точку (газ-жидкость-твердое). По некоторым причинам выделяется как независимое состояние идеального газа.

Для каждого свойства указывается название, сокращенное обозначение и тип. По типу свойства делятся на свойства-функции и свойства-константы. Для свойств-функций определяется набор аргументов (как правило, это температура и/или давление), область определения, область значений, характер монотонности. Числовые значения свойств-функций задаются для определенных значений их аргументов. Свойства-константы не имеют аргументов. Их числовые значения задаются для вещества без указания дополнительных параметров. Свойства могут быть применимы к одним состояниям веществ и не применимы к другим. Числовые значения свойств даются с указанием единиц измерения и сопровождаются указанием на источник/и, причем эта ссылка может относиться к единичному значению, к отдельному свойству или даже ко всему набору, включающему несколько функций.

Между выбранными свойствами и фазовым/агрегатным состоянием, в котором находится вещество существуют жесткие логические связи и ограничения, обусловленные физической природой вещества. Есть свойства, которые по физическому смыслу приложимы только к определенной линии фазового равновесия. Например, такие свойства как энтальпия испарения или температура кипения имеют смысл только для линии насыщения «газ-жидкость». Такое свойство как вязкость имеет смысл только для состояний «газ» или «жидкость». С другой стороны, если в наборе присутствуют данные для обеих сосуществующих фаз, необходимо различать принадлежность свойства веществу в том или ином состоянии, например для газа или жидкости. Эти ограничения должны быть явно отражены в структуре классов и связей. В предыдущих работах [1, 2] были введены основные классы и связи, необходимые для адекватной формализации предметной области. Эти же вопросы, но в применении к задаче публикации данных в LOD были подробно рассмотрены в дипломной работе [3].

Хотя в сети Интернет представлено множество онтологий, включающих понятия и концепции, относящиеся к домену «теплофизические свойства веществ», однако в свободном доступе, по-видимому, нельзя указать онтологию, которая позволила бы в полном объеме представить рассматриваемые данные в RDF. В данной работе, на основе результатов [3] предложена онтология с заимствованием терминов из онтологий **QUDT** (область: численные свойства и их значения, а также единицы измерения) и **ChemAxiom** (область: данные о веществах). Исходные данные набора, представленные в реляционном виде, преобразованы в формат RDF с помощью инструмента D2R.

Процедура включения собственного набора данных в пространство LOD в качестве первого шага предполагает их загрузку в реляционную БД, схема которой дает предварительную формализацию предметной области [4]. На следующем шаге составляется онтология, обогащенная заимствованием терминов из родственных онтологий, к которым имеется доступ в сети. При составлении последней, дающей наиболее адекватную формализацию предметной области, должны быть использованы как ранее введенные классы и связи, так и новые, отображающие ограничения, налагаемые спецификой предметной области. Затем, исходные реляционные данные преобразуются в формат RDF, соответствующий разработанной онтологии. На последнем шаге проводится вручную или автоматически генерация связей данного набора с тематически родственными ресурсами, ранее локализованными в LOD.

2. Тематически близкие онтологии и словари

Ключевую роль при публикации в LOD собственных наборов данных играют онтологии и словари, обеспечивающие возможность семантической интеграции. Идеальным решением проблемы было бы использование одной из доступных онтологий, адекватно отражающих специфику предметной области. Невозможность найти онтологию, полностью покрывающую домен, заставляет выделить в понятийном аппарате отдельные категории с доступным онтологическим описанием. В нашем случае среди таких категорий можно выделить: вещества, свойства, единицы измерения и родственные понятия. При этом выбор внешних онтологий целесообразно подчинить простому требованию: используемые термины должны принадлежать возможно меньшему числу онтологий, то есть выбранные онтологии должны покрывать по возможности большую часть рассматриваемой здесь предметной области.

В качестве онтологий, способных передать терминологию, связанную с понятием «вещество», выбрана система онтологий **ChemAxiom** [5]. К ее достоинствам относится возможность использования данных из БД **ChemSpider** [6], наиболее подходящего источника для идентификации

веществ и связывания с внешними наборами данных. Разработчики **ChemAxiom** [5] создавали общую онтологию химического домена, способную заменить узко-ориентированные онтологии. Система представляет собой набор онтологий-модулей, каждая из которых описывает отдельную подобласть химии. Модули являются автономными, но связаны через онтологию высшего уровня **Basic Formal Ontology**. В главном модуле выделяются понятия «Вещество», «Молекула» и связи между ними. В отдельные классы выделены химические дескрипторы (формула, название и др.). Существует также специальный модуль, определяющий свойства веществ в виде словаря классов без связей с единицами измерения. В этом модуле введены также термины для определения состояний вещества, что подчеркивает близость покрываемой **ChemAxiom** тематики предметной области, связанной с макроскопическими свойствами вещества.

Наряду с онтологией базовых химических понятий, отразить специфику предметной области призван терминологический словарь названий веществ. В качестве последнего принята БД **ChemSpider** [6], одна из наиболее масштабных БД этого профиля. **ChemSpider** содержит данные для более, чем 30 млн. соединений примерно из 400 источников записей, используется как платформа для аннотирования и сопровождения существующих данных, а также как полезный источник сведений о веществе, особенно при отображении его названий. В частности, БД позволяет провести отображение между **ChemSpider identifiers (CSIDs)** и оригинальным источником данных. Например, при поиске в БД вещества под названием “hydrogen” БД выдает его уникальный номер и соответствующий URI, CSID:762, (www.chemspider.com/Chemical-Structure.762.html). По запросу **ChemSpider** выдает подробную запись, включая перечень названий-синонимов, данные о свойствах, источниках и т.п. Главной сущностью является «Вещество», причем для описания данных используется онтология **ChemAxiom**. SPARQL-точки у этого набора пока нет, но функционирует точка поиска, работающая по протоколу HTTP, на которую можно посылать поисковые запросы с выдачей либо RDF-документа, если под критерии поиска подходит одно вещество, либо XML-данные со списком параметров найденных веществ, если таких веществ несколько.

Для представления понятий, связанных со свойствами вещества, принята популярная онтология **QUDT** - Quantities, Units, Dimensions and Data Types in OWL and XML [7]. Семантика **QUDT** основана на анализе размерностей, записанном на языке OWL. Здесь же конкретизируется семантика величин и единиц измерения. Спецификация QUDT обеспечивает интероперабельность и обмен данными за счет доступности в машинно-читаемой форме. Онтология **QUDT** оказалась самой богатой среди аналогичных в плане предоставляемых терминов. Она определяет классы для единиц измерения, свойств, для величин и их значений, определяет словари типов свойств и единиц измерения, в то время, как другие онтологии из рассмотренных в работе [3] определяют средства представления лишь для отдельных понятий предметной области. В частности, другие онтологии не предоставляют возможности работы с размерностями. Онтология **QUDT** предоставляет классы для представления величин и их значений, предикаты для связи экземпляров этих классов, и в то же время, списки типов свойств и единиц измерения. Данная онтология покрывает заданную область в наибольшей степени.

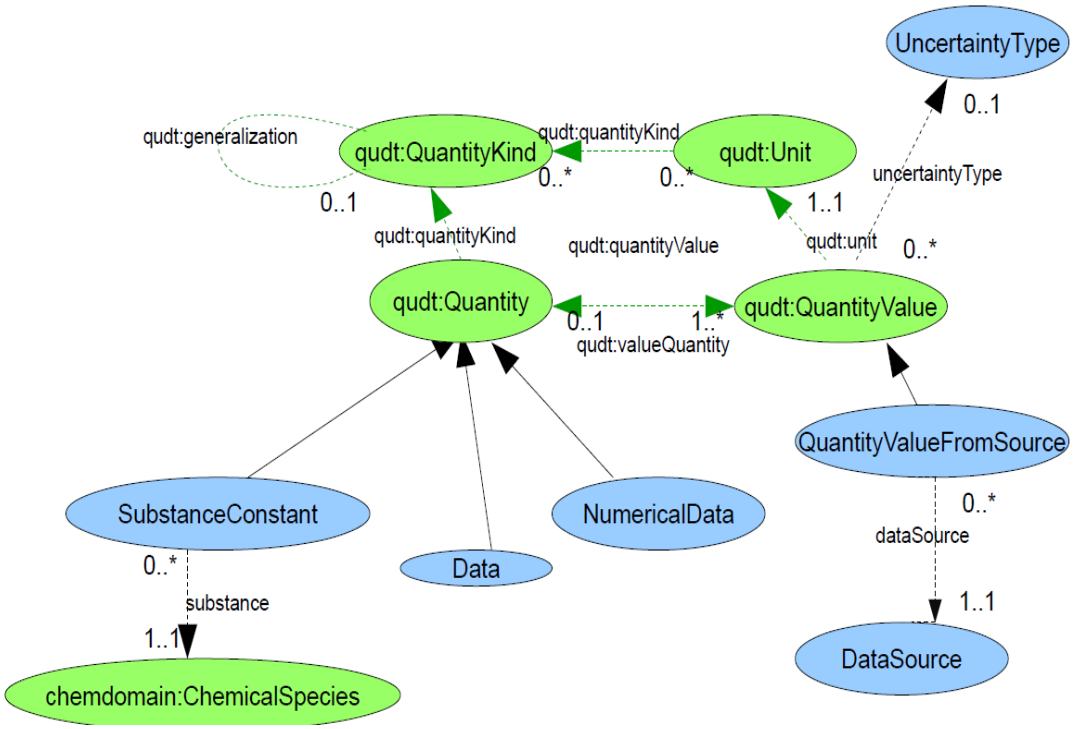
3. Расширенная онтология предметной области

Предыдущий этап включал создание онтологии, на основе которой была предложена схема описания реляционной БД [2]. Здесь на основе этой схемы, достаточно конкретно отражающей специфику предметной области, предложена онтология более высокого уровня, решающая задачи повторного использования терминов из существующих онтологий, а также проверки связей и ограничений между множеством понятий. Поскольку данные исходного (авторского) набора представлены в виде записей реляционной БД, их требуется конвертировать в RDF формат, согласованный с разрабатываемой онтологией. Поэтому для того, чтобы составить требуемую онтологию, нужно, пользуясь схемой реляционной БД, построить соответствие таблицам из БД классов из предметной области. В частности, каждая запись в таблице «substance» преобразована в три RDF ресурса классов ChemicalSpecies, MolecularEntity и BruttoFormula, связанных соответствующими предикатами.

Элементам схемы БД, относящимся к единицам измерения, свойствам, численным значениям ставятся в соответствие термины из онтологии **QUDT**. Таблицам «property» и «dimension», которые хранят информацию об измеряемых свойствах и их единицах измерения, ставятся в соответствие классы QuantityKind и Unit. Схожие типы сущности для предметной области представлены в таблицах «numerical_data» (численные значения аргументов), «constants_of_substance» (численные значения констант) и «data» (численные значения свойств-функций). Общие поля этих таблиц, «property_id», «value», «uncertainty_id» хранят ссылки на свойство, его численное значение и тип неопределенности. Сущности, информация о которых хранится в указанных таблицах БД, можно разбить на пары экземпляров классов qudt:Quantity (величина) и qudt:QuantityValue (значение), связанные предикатом (qudt:quantityValue). Сущность Quantity будет обладать ссылкой (предикат qudt:quantityKind) на измеряемое

свойство (ссылка property_id), а также ссылкой на сущность типа QuantityValue, у которой будет параметр - численное значение (предикат qudt:numericValue). Для обозначения параметров, не предусмотренных **QUDT,** введены собственные предикаты. Например, для связи класса QuantityValueFromSource, обозначающего значение величины из литературы с экземпляром класса DataSource введен предикат dataSource. Введены также классы Data, SubstanceConstant и NumericalData, унаследованные от класса qudt:Quantity. Специальный предикат uncertaintyType дает ссылки на экземпляры класса UncertaintyType, а предикат uncertaintyValue - на численное значение погрешности указанного класса. Укороченная схема построенной онтологии представлена на рис. 1, где выделены классы и предикаты, заимствованные из сторонних онтологий.

Рис. 1. Укороченная схема онтологии



На схеме не показаны два класса Dataset и EnvironmentalConditions, посредством которых задается соответствие значений свойств-функций значениям свойств-аргументов. В итоге, разработанная онтология более подробно описывает предметную область, чем исходная онтология [2] и соответственно схема исходной БД. Были использованы некоторые классы и предикаты из сторонних онтологий (**QUDT**, **ChemAxiom**) и определены виды преобразования, которые могут потребоваться для представления исходных данных к виду, соответствующему полученной онтологии.

4. Преобразование исходных данных в RDF; процедуры связывания с внешними источниками

Для конверсии данных из реляционной БД к RDF формату выбран D2R сервер, как наиболее простое некоммерческое средство, поддерживающее *дереференсирование* HTTP URI ресурсов [8]. Данный инструмент поддерживает работу со многими коммерческими и свободно распространяемыми БД, в том числе с PostgreSQL, использованной в данной работе. Отображение схемы реляционной БД на RDF-модель задается посредством, так называемого, Mapping File. В нем заложена схема соответствий каждой из таблиц БД нескольким классам сущностей, с выборочным определением набора свойств для классов. В состав D2R платформы входит стандартное Web-приложение, которое позволяет просматривать полученные RDF-данные в браузере при удобной навигации по URL ресурсам. Доступ к данным предоставлен по HTTP (просмотр данных-троек ресурса по его URL), а также с помощью SPARQL протокола. Преобразование данных происходит «на лету», то есть взаимодействие происходит непосредственно с данными из реляционной БД без использования промежуточных хранилищ. Тем самым D2R позволяет провести все преобразования исходных данных, необходимые для их публикации в RDF формате и связывания с родственными ресурсами в LOD.

Для эффективной работы по конверсии реляционных данных необходимо в исходной БД предусмотреть дополнительные таблицы для хранения связей (рис. 2).

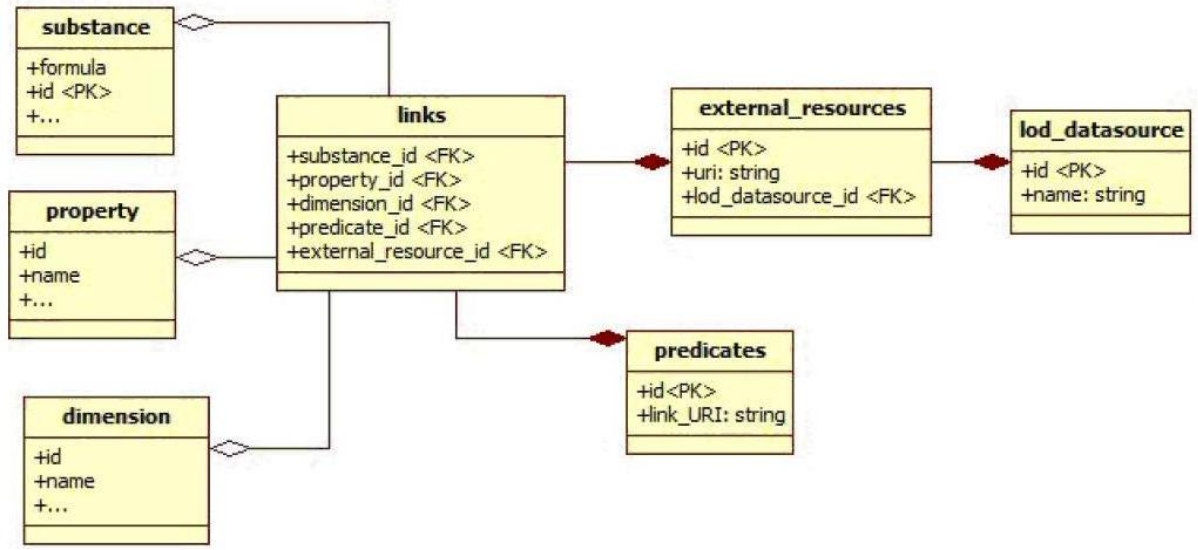


Рис. 2. Дополнительные таблицы для хранения связей

Таблицы «substance», «property» и «dimension» в исходной БД хранят сущности, которые предстоит связать с внешними ресурсами. Остальные таблицы введены специально для организации связей. В таблице «links» указано, сущность какого типа участвует в конкретной связи (первые три поля). Поле «predicate_id» указывает предикат для связи из списка в таблице «predicates». Поле «external_resource_id» указывает, с каким внешним ресурсом производится связывание. Внешние ресурсы хранятся в таблице «external_resources» и могут принадлежать различным источникам данных в LOD («lod_datasource»). Таким образом в каждой записи таблицы «links» хранится тройка **«вещество|единица измерения|свойство – предикат из списка — внешний ресурс»**. При преобразовании с использованием D2R эта запись конвертируется в соответствующую RDF тройку в соответствии с инструкцией, занесенной в Mapping File. Связывание данных проводится по разным процедурам для сущностей-веществ, сущностей-свойств и сущностей-единиц измерения. В задаче связывания сущностей-веществ наиболее эффективно использование БД **ChemSpider**, тем более, что этот ресурс использует онтологию **ChemAxiom**, также выбранную нами для представления сущностей-веществ в публикуемом наборе данных.

Хранение в таблицах БД систематического названия вещества и химической формулы недостаточно для автоматической генерации связей, прежде всего, из-за различий в использованных названиях. Кроме того, стехиометрическая (брутто-) формула вещества не идентифицирует вещество однозначно. Решение этих проблем возлагается на эксперта, который использует пользовательский интерфейс для нахождения соответствия между веществами авторского набора и внешнего ресурса с генерацией соответствующих связей.

Для генерации связей между экземплярами свойств и единиц измерения, определенными в авторском наборе и в онтологии **QUDT** возникают проблемы из-за отсутствуя критериев для автоматической генерации связей и отсутствия в онтологии многих понятий, использованных в авторском наборе. Принятый здесь подход к процессу связывания возлагает ответственность за связывание на пользователя, предоставив ему интерфейс связывания, а также интерфейс определения свойств и единиц измерения в терминах **QUDT**. Далее полученные экземпляры используются как внешние ресурсы для связывания. При этом пользователю предоставлена возможность задавать ограничения на численные значения.

5. Заключение

Итогом работы является созданная технология представления теплофизических данных в пространстве LOD. Основу технологии составляет детализированная онтология предметной области, использующая термины ряда тематически близких онтологий и словарей. Важнейшим элементом онтологии является проверка выполнения различных связей и ограничений между понятиями, обусловленных спецификой предметной области. Построенная онтология по степени отображения предметной области существенным образом расширяет схему описания данных, принятую в реляционной БД для хранения исходных наборов. Конверсия реляционных данных к RDF формату позволяет поддерживать все требования, налагаемые онтологией, и проводить связывание в ресурсами LOD. Наличие открытых списков для важнейших сущностей (вещество, свойство, единица измерения, тип неопределенности)

обеспечивает возможности гибкой подстройки развитой технологии к постоянно расширяемой сфере применения.

Литература

[1] Еркимбаев А.О. и др. Интеграция данных по свойствам веществ и материалов на основе онтологического моделирования предметной области/ Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Серебряков В.А., Шиолашвили Л.Н. // Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012). Санкт-Петербург, 2012. С. 38-47.

[2] Атаева О.М. и др. Интеграция данных по теплофизическим свойствам веществ методами онтологического моделирования (доклад, слайды). / Атаева О.М., Еркимбаев А.О., Зицерман В.Ю., Кобзев Г.А., Серебряков В.А., Теймуразов К.Б., Хайруллин Р.И. // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. XV Всероссийская научная конференция. Ярославль, Россия, 4-17 октября 2013 года. – Ярославль: ЯрГУ, 2013.- 422 с. URL: http://rcdl2013.uniyar.ac.ru/doc/full_text/rcdl_ataeva_i_dr.pdf (дата обращения 08 сентября 2014)

[3] Устинова Е.С. Дипломная работа «Интеграция данных по свойствам веществ в специализированное пространство связанных данных».// МГУ им. М.В. Ломоносова. Факультет вычислительной математики и кибернетики. Кафедра системного программ-мирования. М: 2014.

[4] Health Care and Life Science (HCLS) Linked Data Guide. URL: www.w3.org/2001/sw/hcls/notes/hcls-rdf-guide/ (дата обращения 08 сентября 2014)

[5] Adams N., Cannon E.O., Murray-Rust P. ChemAxiom – An ontological framework for chemistry in Science // Nature Precedings. 2009. URL: http://precedings.nature.com/documents/3714/version/1 (дата обращения 08 сентября 2014)

[6] ChemSpider. Search and share chemistry. URL: www.chemspider.com (дата обращения 08 сентября 2014)

[7] Hodgson R. et al. QUDT - Quantities, Units, Dimensions and Data Types Ontologies. March 18, 2014 /Hodgson R., Keller P.J., Hodges J., Spivak J. URL: http://qudt.org/ (дата обращения 08 сентября 2014)

[8] Sahoo S.S. et al. A Survey of Current Approaches for Mapping of Relational Databases to RDF. //W3C RDB2RDF Incubator Group, January 08 2009. URL: http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf (дата обращения 08 сентября 2014)

Sharing Thermophysical Properties data on the LOD (Linked Open Data)

V.A. Serebriakov, K.B. Teimurazov, E.I. Ustinova, A.O. Erkimbaev, G.A. Kobzev, V.Yu. Zitserman

Detailed ontology for thermophysical data linking on the LOD is developed. Provision is made for terms from external ontology (ChemAxiom, QUDT) reuse along with inserting the linkage between main concepts. Original data presented as relational data base records are mapping to RDF format by the use of the shareware D2R server. It is developed the technology for linkage of the author data set with the LOD resources which are related to substances, properties and units of measurements.

♣ Работа выполнена при поддержке РФФИ, грант №13-07-00218.