

Онтология как средство интеграции автономных баз данных по свойствам наноматериалов

А.В. Косинов¹, А.О. Еркимбаев¹, Г.А. Кобзев¹, В.Ю. Зицерман¹

¹Объединенный институт высоких температур РАН, Ижорская 13, стр. 2, Москва, Россия, 125412

Аннотация. Доклад касается проблем, относящихся к новой дисциплине под названием наноинформатика, предметом которой является разработка методов систематизации и распространения данных и знаний по свойствам наноматериалов и основанных на них устройствах и технологиях (Еркимбаев и др. *Наноинформатика: задачи, методы и технологии. НТИ, сер. 1. 2016(10), 1-18*). Ведущим направлением в этой дисциплине является интеграция разнородных ресурсов в виде баз данных и Web-порталов, с существенными различиями в содержании, структуре и форматах данных. В сфере нанотехнологий на пути интеграции данных возникает целый ряд барьеров, связанных с междисциплинарным характером предметной области, неустойчивостью терминологии и понятийного аппарата, перманентным появлением новых материалов и устройств. В докладе показано, что ключом к решению этих проблем является использование онтологий как эффективного средства стандартизации данных и знаний. Предложена онтология наноматериалов, включающая словарь базовых понятий и концептуальную схему, формирующую логические связи между ними. Разработаны программные средства для управляющей надстройки, которая оказалась способной интегрировать ранее созданные автономные базы данных по углеродным наноматериалам и наноматериалам для энергетики.

1. Введение

Понятие об онтологии уже давно заняло одно из ведущих мест среди информационных средств организации и распространения научных знаний. Онтология включает единый, признанный в профессиональном сообществе словарь всех понятий, расширенный за счет логических связей и аксиом. В закодированном виде онтология, описывая знание с возможностью его машинной интерпретации, становится управляющей надстройкой, способной интегрировать разнородные данные. Важнейшая особенность онтологии – возможность ее перманентного расширения, необходимая по мере введения в практику новых материалов и связанных с этим понятий. Среди ее возможных использований – стандартизация профессиональной лексики, исключая неоднозначность и ложную интерпретацию, связывание документов с родственным контентом методами Semantic Web; формирование структуры данных и знаний для распространения в сети.

Осознание потенциала онтологии для интеграции и распространения научно-технических знаний стимулировало активную деятельность по разработке их предметно-ориентированных версий для формализации отдельных сегментов многих научных дисциплин: химии, материаловедения, наук о Земле и др. [1].

Естественно, что сфера нанотехнологий, как молодая междисциплинарная область, породившая беспрецедентный по объему поток данных в виде публикаций и патентов, также не осталась в стороне от общей тенденции. Из разработанных и активно применяемых онтологий можно отметить NPO (NanoParticle ontology for cancer nanotechnology research) [2], онтологию, встроенную в химическую базу данных (БД) ChEBI [3], ориентированную на свойства распространенных химических веществ и молекулярных сущностей (частицы, молекулярные фрагменты, комплексы), онтологию мета-уровня, предложенную Tanaka [4], которая представляет собой обширный междисциплинарный классификатор, таксономию на портале InterNano [5], и некоторые др. Подробный обзор и анализ их возможностей дан в обзоре [6].

Важнейшим направлением практического использования онтологий стало в последние годы создание на их основе управляющей надстройки над группой разнородных и пространственно распределенных ресурсов с целью их интеграции, то есть виртуального объединения с общей точкой доступа, интерфейсом, и системой запросов [7]. Недавно разработанный нами сценарий [8] предусматривает совместное применение онтологических моделей и инструментария технологий Big Data (Spark SQL) для задач интеграции и оперативной работы с разнородными по типу данными о свойствах веществ и материалов. Сценарий предполагает конверсию множества исходных ресурсов (БД, файловых архивов и т.п.) к стандартному текстовому формату JSON с последующим использованием онтологий для семантической интеграции. Ключевой элемент технологии заключается в связывании атрибутов (ключей), а также и по возможности значений (объектов) с концептами онтологий.

Потребность в обращении к инструментарию Big Data связана не столько с объемом данных, сколько с многообразием условий синтеза и измерений, микроструктурных особенностей и проч. Платформа Apache Spark (<https://spark.apache.org/>) и встроенная в нее библиотека Spark SQL оказались наилучшим образом приспособлены к работе с данными разной структуры: реляционными, неструктурированными, файловыми архивами и др. Интеграцию данных обеспечивает приведение их атрибутов к единым концептам предметно-ориентированной онтологической модели. Применение файлов в JSON формате позволяет организовать выгрузку данных с совершенно разной структурой практически из всех типов БД, не прибегая к прямому подключению.

Наконец, использование программных средств Big Data, в частности Spark SQL, дает возможности производительного оперирования наборами файлов в JSON формате. При этом манипуляции на уровне концептов/терминов происходят на языке запросов к онтологиям SPARQL (рекурсивный акроним от англ. SPARQL Protocol and RDF Query Language), а запросы к содержательной части на языке SQL (Structured Query Language). При этом использование онтологий дает значительный простор для решения вопросов, связанных с изменчивостью структуры данных, что почти невозможно при эксплуатации традиционных БД.

В данной работе предложенная технология использована для интеграции ранее созданных БД по свойствам наноматериалов (НМ). В качестве исходных ресурсов приняты две тематически родственные БД, из которых одна охватывала данные по свойствам углеродных НМ различного генезиса и применения, а вторая – произвольные по химическому составу НМ, объединенные одним признаком – потенциалу использования в сфере энергетики. Отличаются они и по структуре данных. Первая построена как хранилище полуструктурированных данных с вариацией числа и типа атрибутов в зависимости от класса НМ, вторая – как БД с жесткой структурой.

2. Первичный вариант онтологии по наноматериалам

Для решения поставленной задачи – интеграции тематически родственных БД по НМ, в работе использована, так называемая, «облегченная» онтология (*lightweight ontology*), то есть содержащая лишь основные структурные и лексические элементы предметной области. При этом, разработка, развитие и модернизация онтологии (усложнение или упрощение) это естественный рабочий процесс для предложенной технологии [8]. На рисунке 1 показана иерархия ее ведущих классов, записанная посредством редактора Protégé [9]. Построение онтологии [10] для научных дисциплин оправдано начинать с выделения фундаментальных

понятий, заимствованных из предметно-независимой онтологии **BFO** (basic formal ontology). В этой схеме исходный класс **Entity** имеет два наследника: **Continuant** для сущностей, сохраняющих со временем свою идентичность, и **Occurent**, для сущностей, предположительно меняющихся со временем. В свою очередь, класс **Continuant** имеет два наследника - **IndependentContinuant** и **DependentContinuant**, охватывающих объекты, относящиеся к предметной области, и их атрибуты. Что касается класса **DependentContinuant**, то он также имеет два дочерних класса: **SpecificallyDependentContinuant** (характеристики объектов) и **GenericallyDependentContinuant** (суммирующий все виды данных и требуемых для этого концепций) с одним дочерним классом **InformationContentEntity**. Класс **SpecificallyDependentContinuant** имеет два дочерних класса: **Quality** (совокупность свойств) **Role**, определяющий объекты в терминах их возможных ролей (функций, приложений) или способностей (disposition). Для упрощения на рисунке 1 показан класс **AreasOfApplication**, по сути являющийся дочерним классом по отношению к **Role**. В итоге формируется строгая и универсальная структура, легко адаптируемая к предметной области.

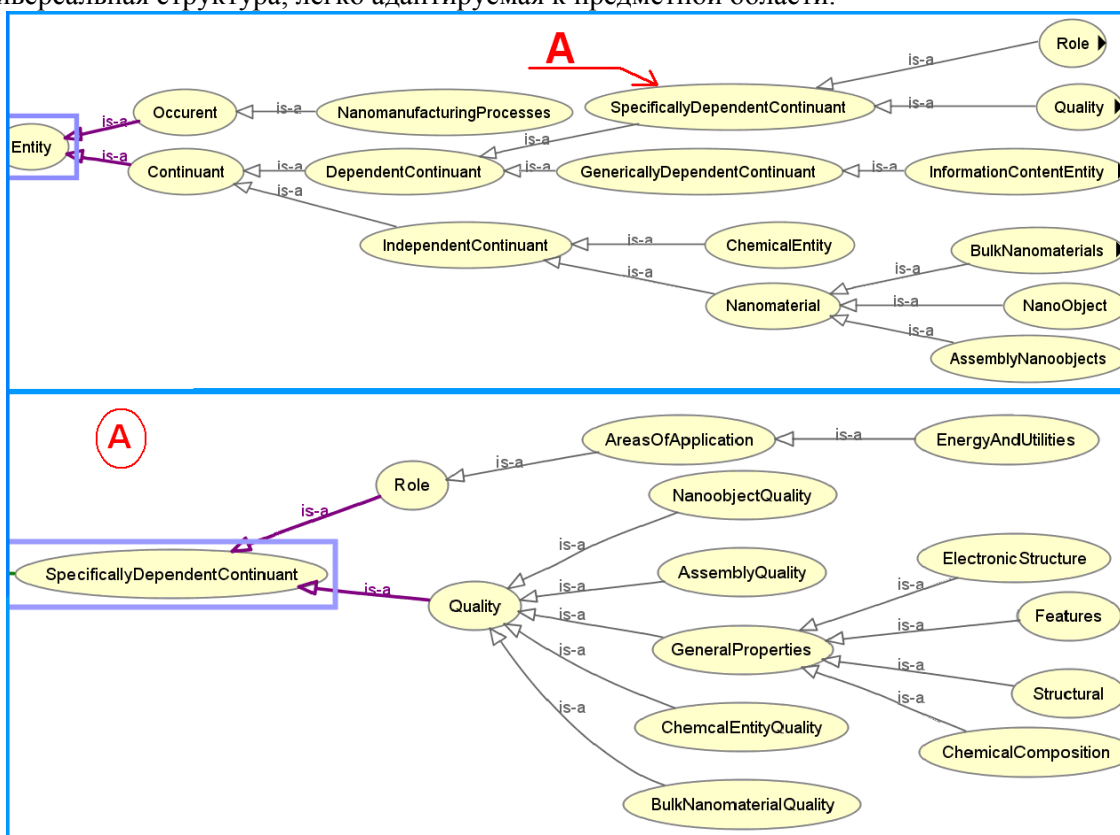


Рисунок 1. Упрощенная версия онтологии НМ.

Ее специфика, определяемая, прежде всего, выбором материальных объектов, представлена классами, дочерними относительно **IndependentContinuant**. В соответствии со стандартом описания НМ, разработанным CODATA [11], все множество НМ, разбивается на три категории: нано-объекты, их ансамбли (коллекции) и макроскопические (или объемные) НМ с присущей им внутренней или поверхностной наноструктурой. Критерием для выделения нано-объекта служит наличие в его геометрии хотя бы одного масштаба в наномасштабе (например, 1 масштаб для графена и 2 масштаба для нанотрубки). К ансамблю нано-объектов относят, согласно [11], структуру, образованную агрегированием за счет химических или ван-дер-ваальсовых связей нано-объектов нескольких типов, например графена и нанотрубок. Как видно из рисунка 1, каждому виду НМ в онтологии отвечает определенный класс, дочерний относительно **IndependentContinuant**. На том же уровне представлен класс **ChemicalEntity** для

молекул и веществ, определяющих состав НМ и/или включенных в процессы изготовления и использования.

Класс **InformationContentEntity** суммирует все сущности, связанные с представлением данных, прежде всего, с необходимостью именования и классификации НМ. К ним относится, в частности, **IdentifyingDescriptor** включающий набор принятых в литературе классификаторов, ссылка на чьи рубрики позволяет четко выделить группу объектов или процессов. Характерным примером служит таксономия **InterNano** [5], отражающая не только вид НМ, но его применение, способ производства, контроля и т.п. Каждая из рубрик имеет соответствующий 5-значный индекс, включенный в состав URL. Например, URL для рубрики **AreasOfApplication** имеет вид <http://www.internano.org/taxonomy/term/11945>, а для рубрики следующего уровня **EnergyAndUtilities** - <http://www.internano.org/taxonomy/term/17875>. Использование ссылок на индекс **InterNano** позволяет определить всю иерархию классов, дочерних относительно **AreasOfApplication** и **NanomanufacturingProcesses**. Другой классификатор, принятый в нашей работе [12], позволяет достаточно детально определить виды НМ, основываясь на их топологии и сфере использования. Ссылки на него суммированы в классе **OriginalDescriptor**.

Наконец, последняя и наиболее значимая часть иерархии охватывает все понятия, раскрывающие физические и эксплуатационные свойства НМ. Вообще говоря, перечень этих характеристик весьма обширен и существенно зависит как от вида НМ, так и сферы его использования. В данной работе осознанно принята крайне урезанная иерархия классов, достаточная для иллюстрации процедуры интеграции БД. Каждому из трех видов НМ поставлен в соответствие класс, отражающий характерные только для него свойства: **NanoobjectQuality**, **AssemblyQuality**, **BulkNanomaterialQuality**. При необходимости для каждого из них следует развернуть свою иерархию. Так для определения свойств ансамбля должен быть детализирован его состав с указанием вида и числа нано-объектов, наподобие молекулярной формулы. Подробные перечни свойств, необходимых для характеристики нано-объектов и их ансамблей, можно найти в рекомендациях CODATA [11]. Построение онтологии на их основе, причем с обязательным охватом макроскопических НМ, обсуждалось в работе [6]. Здесь для этих целей введен класс **GeneralProperties**, объединяющий те свойства, что могут быть отнесены к любому из классов НМ. Классы, приведенные на рисунке 1, охватывают лишь небольшую группу характеристик: химический состав, структура, включая кристаллографические данные, электронная структура, например ширина запрещенной зоны. При необходимости этот перечень должен быть расширен за счет тепловых, механических, оптических и др. физических и эксплуатационных свойств. Особую роль при характеристике НМ играет класс **Features**, который суммирует априори непредсказуемые физические факторы, определяющие специфику материала. Они могут включать состояние поверхности, морфологию образца, дефекты структуры, наличие примесей и проч. Наличие таких факторов, существенно влияющих на свойства, является характерной особенностью всех НМ, хотя отчасти проявляется и для традиционных материалов типа сталей или керамики. Набор классов, дочерних относительно класса **Features**, всегда зависит от сегмента предметной области и характерного вида НМ.

3. Применение технологии интеграция для автономных баз данных по НМ

Технология, предложенная нами ранее [8], позволяет, основываясь на онтологии или контролируемом словаре, обеспечить интеграцию разнородных ресурсов, отличающихся словарем понятий, структурой и форматом данных. Ключевой элемент технологии – процедура связывания классов онтологии с атрибутами, принятыми в том или ином ресурсе, например в БД. Очевидно, что все этапы, как построение онтологии, так и связывание ее с разнородными ресурсами не может происходить без участия эксперта в предметной области, который только и способен определить близость и логические отношения между понятиями/метаданными.

Ниже в схематичном виде показано, как реализуется эта процедура на примере двух ранее созданных БД, различающихся принципом выбора НМ. Первая из БД предназначалась для охвата всевозможных наноструктур углерода, с детализацией их свойств, метода изготовления

и потенциала использования [12]. Выбор углерода был обусловлен беспрецедентным многообразием его наночастиц, различающихся по размеру, морфологии, характеру связей и т.п., и уникальными свойствами уже изученных объектов (фуллеренов, нанотрубок, графена и др.), перспективных для новых технологий и устройств. При построении БД учитывалась основная особенность предметной области – постоянные вариации структуры данных за счет многообразия свойств и объектов, так называемая полуструктурированность данных. Так, например, фуллерены типа C_{60}, C_{70} характеризуют по аналогии с обычными многоатомными молекулами, для углеродных нанотрубок используют специфический параметр, под названием хиральность [13], а для макроскопических объектов типа фуллеритов данные о кристаллической структуре вещества. Помимо этого, для НМ в еще большей степени, чем для обычных (сталь, пластик, керамика) проявляется влияние множества экстра-факторов: размеры и конфигурация образца, его предыстория, внешние поля, состояние поверхности и проч. В упрощенном виде принятая структура данных представлена на рисунке 2. На верхнем уровне задан «жесткий скелет» из 4 блоков (**Source, Identification, Properties, Application**), из которых наиболее сложным с точки зрения состава и структуры оказывается второй, чьей задачей является однозначная идентификация НМ, рисунок 3.

Source	Identification	Properties	Application
<i>Bibliographic data</i>	<i>Set of identifying signs</i>	<i>Physical and performance properties – data and metadata. Uncertainty estimations.</i>	<i>Area of application</i>

Рисунок 2. Верхний уровень логической структуры данных для представления углеродных НМ.

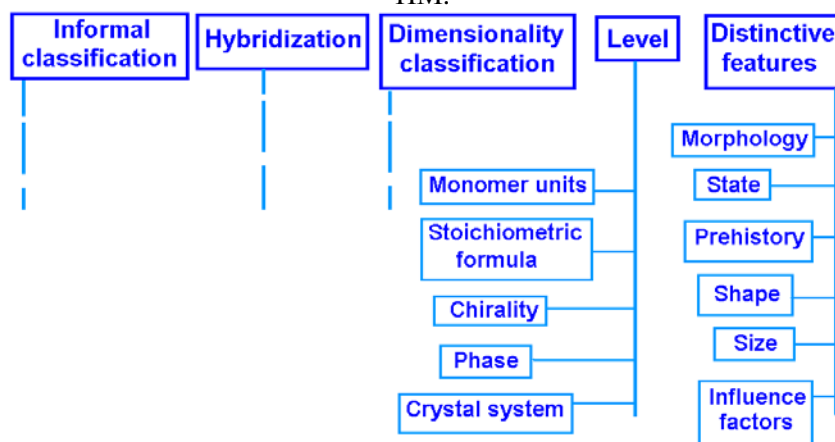


Рисунок 3. Типовая структура блока данных для идентификации углеродных НМ.

Три его первых атрибута (**Informal classification, Dimensionality classification, Hybridization**) дают общее представление о характере объекта, не выделяя его однозначно из класса однотипных. Под термином **Informal classification** понимается общеупотребительное название класса объектов, например фуллерены, нанотрубки, графен и т.п. Второй тип классификации, принятый в работе [14], основывается на числе измерений макроскопического масштаба, как для самого НМ, так и составляющих его фрагментов. Наконец параметр под названием **Hybridization**, рекомендованный многими авторами в качестве основы классификации углеродных материалов [15, 16], дает указание на тип химической связи (sp^1, sp^2, sp^3 и/или их комбинация). Задание сложной и варьируемой структуры данных возложено на два последних идентификатора (**Level, Distinctive features**), рисунок 3. В наших работах использовалась БД PostgreSQL (<https://www.postgresql.org/>), в которой для этой цели предусмотрен, так называемый, *композиционный* тип данных, объединяющий элементарные типы данных. В его

рамках удастся отстроить иерархию для сколь угодно глубокой детализации объекта, а варьируя число и тип атрибутов добиться требуемой изменчивости структуры данных.

Атрибут **Level** производит разделение объектов **макро-**, **микро-** (одиночные молекулы) и **мезо-** уровней (промежуточные по масштабам и природе). В зависимости от этого привлекаются различные атрибуты, как то: мономерная единица, стехиометрическая формула, параметр **Chirality**, используемый для нанотрубок и пластин графена, данные о кристаллической структуре и т.п. Другой атрибут, **Distinctive features**, также представленный посредством композитного типа данных, интегрирует все сведения, детализирующие признаки НМ, рисунок 3. В принципе, перечень рубрик на всех уровнях иерархии может быть заметно расширен, если основываться на международных рекомендациях, подготовленных CODATA [11] для целей многофакторной однозначной идентификации НМ.

Блок данных **Properties** (рисунок 2) использует классификатор, включающий типовые классы физико-химических характеристик (термодинамические, структурные, механические и т. п.) Каждая из групп охватывает перечень понятий (метаданных), используемых в литературе по свойствам НМ. Отдельный классификатор позволяет также аттестовать меру достоверности данных, ориентируясь на данные измерений и оценку их полноты и согласованности [13].

Вторая из ранее подготовленных БД по свойствам НМ [17] отличается как принципом отбора данных, так и содержанием записей. Отбирались данные только для тех НМ, которые используются (или имеют потенциал использования) в различных секторах энергетики: солнечной, электрохимической, атомной и др. При этом допускается произвольный химический состав материала, также как и его топология, размеры, форма и прочие характеристики. Отличия затрагивают и структуру данных – принят постоянный перечень метаданных, определяющий сам НМ, его свойства, применение и метод синтеза. Гибкость в представлении данных достигается за счет сочетания классификаторов для выделения признаков из априори заданной схемы и текстовых записей для развернутого определения. Часть полей позволяет записать фактографические данные в формате RTF (rich text format), сочетая фрагменты текста, математические символы и графику для исчерпывающей информации по методам синтеза НМ, его свойствам (физическим и эксплуатационным) и сфере приложения. Любые термины, используемые в этих полях, могут использоваться при полнотекстовом поиске. БД позволяет также включить в запись произвольное число гиперссылок для обращения к файлам текстовых документов на сервере и к удаленным серверам, где находятся тематически родственные документы, включая электронные справочники, БД и т.п.

Несмотря на существенные различия обеих БД (по углеродным НМ и НМ для энергетики) в отношении структуры данных, между ними имеется и заметное пересечение содержания, что делает актуальным создание управляющей надстройки, обеспечивающей их интеграцию в рамках единой онтологии. Разработанная процедура [8] состоит в предварительной конверсии каждой из БД к документу в JSON-формате с последующим связыванием их метаданных с единой онтологией. Наиболее просто эта процедура реализуется для второй БД, где принят фиксированный набор метаданных, таблица 1.

В онтологии для каждого из них можно подобрать адекватный класс, либо идентичный по смыслу, либо покрывающий более широкое понятие; в последнем случае метаданные могут трактоваться как суб-классы по отношению к соответствующему классу из онтологии. Например, понятие **Energy sector** может рассматриваться как эквивалент классу Energy and Utilities, а более узкое понятие **Energy function** как его суб-класс. В БД для определения НМ использовался оригинальный классификатор, наличие которого в онтологии отмечено классом **Original_descriptor**. Для детализации метода синтеза и сферы приложения в онтологии импортированы термины из **InterNano taxonomy** [5], каждый из которых индексирован соответствующим URL и номером.

Сложнее реализуется связывание онтологии с атрибутами углеродной БД из-за принципа ее построения как хранилища *полуструктурированных* данных. За этот счет резко нарастает число метаданных, определяющих номенклатуру требуемых характеристик. По мере появления новых данных, возникают ранее не предусмотренные понятия, требуя новых операций

связывания, а возможно, и расширения онтологии путем введения новых классов. Здесь для примера рассмотрено связывание с онтологией блока данных **Identification**, заполненного данными для НМ под названием **Peapod** (нанотрубка, заполненная несколькими молекулами фуллерена). Пять фиксированных метаданных верхнего уровня и их связи с классами онтологии выделены жирным шрифтом в строках таблица 2.

Таблица 1. Метаданные БД по НМ для энергетики [17] и их связывание с классами онтологии.

Метаданные	Классы онтологии	Область применения метаданных
Nanomaterial [free title]	Common_name	Эквивалентна
Nanomaterial by rubricator	Original_descriptor	Эквивалентна
Chemical [free title]	Chemical composition	Суб-класс
Chemical by rubricator	Chemical composition	Суб-класс
Synthesis	Nanomanufacturing Processes	Эквивалентна
Properties	Quality	Эквивалентна
Application	Areas of Application	Эквивалентна
Energy sector	Energy and Utilities	Эквивалентна
Energy function	Energy and Utilities	Суб-класс
Object	Energy and Utilities	Суб-класс

Таблица 2. Метаданные БД по углеродным НМ и их связывание с классами онтологии.

Метаданные	Классы онтологии	Область применения метаданных
Informal classification	Common name	Эквивалентна
Dimensionality classification	Structural	Суб-класс
Hybridization	Electronic_Structure	Суб-класс
Level	Nanomaterial	Эквивалентна
Monomer units	Assembly_quality	Суб-класс
Stoichiometric formula	Assembly_quality	Эквивалентна
Chirality	Assembly_quality	Суб-класс
Crystal system	Crystal structure	Эквивалентна
Distinctive_features	Features	Эквивалентна
Morphology	Features	Суб-класс
State	Features	Суб-класс
Prehistory	Occurent	Суб-класс
Shape	Shape	Эквивалентна
Size	Size	Эквивалентна
Influence factors	Features	Суб-класс
Properties	Quality	Эквивалентна
Application	Areas of Application	Эквивалентна

Два из них, определяющих НМ по его топологии (размерности) и химической связи, не имеют прямого аналога в иерархии классов и при отнесении к некоторому классу должны рассматриваться на правах суб-классов. Так, принятая в [14] **Dimensionality classification** может быть отнесена к классу **Structural** (структурные свойства), а термин **Hybridization** отнесен к классу **ElectronicStructure**. Вообще говоря, отсутствие дочернего класса суб-класса в онтологии не требует ее обязательного расширения. Просмотр классов онтологии посредством SPARQL обеспечит нахождение всех связанных с классами терминов и при более узкой области их определения.

Содержание двух понятий, именованных в БД терминами **Level** и **DistinctiveFeatures**, раскрывается привлечением подчиненных понятий – в таблице 2 они записаны обычным шрифтом. Выбор этих понятий (метаданных) полностью зависит от вида НМ в соответствии с принципом формирования полуструктурированных данных. Поскольку НМ **Peapod** относится к классу **AssemblyNanoobjects**, понятия, детализирующие термин **Level**, по большей части должны быть связаны с классом **AssemblyQuality**, который порождает целую иерархию классов, раскрывающих топологию и состав элементов ансамбля.

Остальные термины соотнесены с суб-классами относительно класса **GeneralProperties**, суммирующего все характеристики, которые могут быть приписаны любому из трех видов НМ, рисунок 1. Из этих классов наибольшее число понятий охватывает класс **Features**. Именно с ним связаны типовые характеристики НМ: **Morphology**, **State**, **Influence factors**. Два других понятия (**Shape**, **Size**) присутствуют в качестве эквивалентных классов онтологии, а такой фактор как **Prehistory** (предыстория в состоянии образца или его подготовка) должен быть приписан суб-классу относительно класса верхнего уровня **Occurent**.

Первая фаза в работе пользователя состоит в поиске адекватных метаданных из обеих БД путем SPARQL запроса к онтологии, см. раздел 1. На второй фазе пользователь средствами Apache Spark формирует SQL запрос к фактографическим данным. Простейший пример - обращение к классу **Common_name**, который в обеих БД связан с метаданными, предусматривающими свободное именование любого вида НМ. При более сложном поиске данных для композитной структуры (образованной из разных наноструктур, наподобие **Peapod**) обращение к классу **Assembly_quality** выделяет адекватные метаданные из углеродной БД, а для привлечения второй БД требуется класс **Original_descriptor**, эквивалентный соответствующему рубриктору, таблица 1. В целом, технология интеграции позволяет сосредоточиться поиском данных на верхнем уровне, не вникая в детали концептуальной схемы каждой из объединяемых БД.

4. Заключение

Поскольку выбор в БД подчиненных понятий зависит от вида НМ и может различаться для разных записей, процедура их связывания с онтологией заканчивается лишь по исчерпанию всех записей и, вообще говоря, должна быть возобновлена при расширении БД. Относительная громоздкость процедуры является платой за использование полуструктурированных данных. Вся процедура оказалась бы заметно проще, если бы построение БД сразу проводилось на основе достаточно развернутой онтологии, что обеспечило бы и вариации структуры данных, и их жесткую связь с классами онтологии. В то же время, связывание метаданных с классами онтологии оказывается очень удобным применительно к обычным реляционным БД, как видно из примера с БД по НМ для энергетики (таблица 2), а также применительно к другим типам информационных ресурсов с фиксированной концептуальной схемой.

5. Литература

- [1] Brodaric, B. *Ontology use for semantic e-Science* / B. Brodaric, M. Gahegan // *Semantic Web*. – 2010. – Vol. 1. – P. 149-153.
- [2] NPO: NanoParticle Ontology for Cancer Nanotechnology Research [Electronic resource]. – Access mode: www.nano-ontology.org (01.12.2018).
- [3] The database and ontology of Chemical Entities of Biological Interest [Electronic resource]. – Access mode: www.ebi.ac.uk/chebi (01.12.2018).
- [4] Tanaka, M. *Toward a Proposed Ontology for Nanoscience* [Electronic resource]. – Access mode: http://www.cais-acsi.ca/proceedings/2005/tanaka_2005.pdf (01.12.2018).
- [5] InterNano – Resources for Nanomanufacturing National Nanomanufacturing net [Electronic resource]. – Access mode: http://eprints.internano.org/view/internano_taxonomy/ (01.12.2018).
- [6] Еркимаев, А.О. *Наноинформатика: задачи, методы и технологии* / А.О. Еркимаев, В.Ю. Цицерман, Г.А. Кобзев, М.С. Трахтенгерц // *Научно-техническая информация*. – 2016. – № 10. – С. 1-18.

- [7] Еркимбаев, А.О. Связывание онтологий с базами данных по свойствам веществ и материалов /А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев, А.В. Косинов // Научно-техническая информация. – 2015. – № 12. – С. 1-16.
- [8] Erkimbaev, A.O. Standardization of Storage and Retrieval of Semi-structured Thermophysical Data in JSON-documents Associated with the Ontology /A.O. Erkimbaev, V.Yu. Zitserman, G.A. Kobzev, A.V. Kosinov // Proceedings of the XIX International Conference “Data Analytics and Management in Data Intensive Domains” (DAMDID/RCDL), 2017 [Electronic resource]. – Access mode: <http://ceur-ws.org/Vol-2022/paper36.pdf> (01.12.2018).
- [9] A free, open-source ontology editor and framework for building intelligent systems [Electronic resource]. – Access mode: <https://protege.stanford.edu/> (01.12.2018).
- [10] Uschold, M. Ontologies: principles, methods and applications / M. Uschold, M. Gruninger // Knowledge Engineering Review. – 1996. – Vol. 11(2). – P. 93-136.
- [11] Uniform Description System for Materials on the Nanoscale. Prepared by the CODATA-VAMAS Working Group On the Description of Nanomaterials [Electronic resource]. – Access mode: www.codata.org/nanomaterials (01.12.2018).
- [12] Еркимбаев, А.О. Систематизация данных по физико-химическим свойствам и применению углеродных наноструктур / А.О. Еркимбаев, В.Ю. Зицерман, Г.А. Кобзев // Теплофизика высоких температур. – 2010. – Т. 48, № 6. – С. 869-876.
- [13] Eletsii, A.V. Properties of nanostructures: data acquisition, categorization, and evaluation / A.V. Eletsii, A.O. Erkimbaev, G.A. Kobzev, M.S. Trachtengerts, V.Yu. Zitserman // Data Science Journal. – 2012. – Vol. 11. – P. 126-139.
- [14] Pokropivny, V.V. New dimensionality classifications of nanostructures / V.V. Pokropivny, V.V. Skorokhod // Physica E. – 2008. – Vol. 40. – P. 2521-2525.
- [15] Inagaki, M. Nanocarbons / M. Inagaki, L.R. Radovic // Carbon. – 2002. – Vol. 40. – P. 2263-2284.
- [16] Suárez, I. Nomenclature of sp² carbon nanoforms. Guest Editorial / I. Suárez, N. Grobert, C.P. Ewels // Carbon. – 2012. – Vol. 50. – P. 741-747.
- [17] Trachtengerts, M. Using a digital library instead of a traditional data base: ABCD-based infrastructure for nanotechnology / M. Trachtengerts, A. Erkimbaev, V. Zitserman, G. Kobzev // The Electronic Library. – 2015. – Vol. 33(5). – P. 916-927.

Ontology as a integration means of stand-alone databases on the properties of nanomaterials

A.V. Kosinov¹, A.O. Erkimbaev¹, G.A. Kobzev¹, V.Yu. Zitserman¹

¹Joint Institute for High Temperatures of Russian Academy of Sciences, Izhorskaya st. 13, bd. 2, Moscow, Russia, 125412

Abstract. The report is devoted to the problems of nanoinformatics, focused on the acquisition, categorization, and sharing on nanomaterials data and devices and technologies based on them (Erkimbaev et al. Nanoinformatics: Problems, Methods, and Technologies. Scientific and Technical Information Processing. 2016, 43(4), 199). Key task of a new discipline is the integration of diverse data represented in a variety of databases and Web-resources, differing in content, structure and format of data. New barriers to the integration of data on nanomaterials and nanotechnologies arise from cross-disciplinary nature of the domain, the instability of the terminology and conceptual apparatus, the permanent emergence of new materials, devices, and technologies. The report shows that the key to solving these problems lies in the use of ontologies as an effective tool for standardizing data and knowledge. An nanomaterials ontology is proposed, which functions as a vocabulary of basic concepts and a conceptual scheme that forms the logical connections between them. The developed toolkit was used to build a control superstructure that integrates previously created databases on carbon nanomaterials and energy applications of nanotechnologies.