Высокоуровневая формализация предметной области для консолидации информационных ресурсов в области неорганического материаловедения

© В.А. Дударев[1,2] © Н.Н. Киселева[1]
[1]Институт металлургии и материаловедения им. А.А. Байкова РАН,
[2]Национальный исследовательский университет «Высшая школа экономики»,
Москва, Россия

vic@imet.ac.ru kis@imet.ac.ru

Аннотация. Обоснована актуальность интеграции информационных систем по свойствам неорганических веществ и материалов. Отмечено, что консолидация возможна только на основе формализации предметной области. Введены основные определения и предложена формализация содержимого информационных систем по свойствам неорганических веществ и материалов на базе трех моделей: вербальной, теоретико-множественной и объектно-ориентированной.

Ключевые слова: интеграция информационных систем, неорганическая химия.

High-level Formalization of Problem Domain for Inorganic Materials Science Information Resources Consolidation

© V.A. Dudarev[1,2] © N.N. Kiselyova[1]
[1]Institution of Russian Academy of Sciences A.A. Baikov Institute of Metallurgy and Materials Science RAS,
[2]National Research University Higher School of Economics,
Moscow, Russia

vic@imet.ac.ru kis@imet.ac.ru

Abstract. Information systems on inorganic substances and materials properties integration actuality is grounded. It's noted that consolidation is possible on basis of subject domain formalization only. The paper introduces principal terms definitions and proposes high-level formalization of information systems on inorganic substances properties contents by means of three models: verbal, set-theoretical and object-oriented.

Keywords: information system integration, inorganic chemistry.

1 Введение

Современные исследования во многих областях науки отличаются интенсивным накоплением и обработкой больших массивов данных. Развитие неорганической химии, как науки, привело к огромному числу исследовательских работ, направленных на всестороннее исследование свойств различных классов неорганических веществ. Результаты этих исследований, как правило, оформляются в виде текстов научных работ, что на данном этапе развития информационных технологий (ИТ) делает практически невозможным компьютерный анализ и обработку имеющихся публикаций с целью извлечения из них знаний и фактов.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

Разработка специализированных информационных систем (ИС) по свойствам неорганических веществ и материалов (СНВМ) является необходимым для успешного развития многих наукоемких областей современной промышленности, например, электроники и машиностроения, т. к. позволяет выбрать оптимальные материалы для решения возникающих задач. Поэтому во многих развитых странах вкладываются значительные инвестиции в создание и развитие ИС СНВМ и расчетных систем, в том числе, на основе машинного обучения [1], которые являются по сути инфраструктурным фундаментом не только для инновационной промышленности, но и для самой науки о материалах.

2 Трудности доступа к информации по СНВМ

Необходимо отметить, что не существует ИС СНВМ, которая содержала бы все требуемые для

анализа данные, и часто информация распределена по нескольким ИС СНВМ, поэтому на практике доступ к такой распределенной по разнообразным источникам информации и ее всесторонний анализ даже для специалиста являются проблемой, решение которой неизбежно связано с двумя задачами. Во-первых, для поиска необходимой информации требуется знать, как минимум, перечень ИС СНВМ, в которых может содержаться искомая информация. Во-вторых, специалисту необходимо, имея доступ к целевым ИС, осуществить поиск необходимой информации и ее всесторонний анализ.

Решение первой задачи поиска нужной ИС СНВМ облегчается за счет использования специализированной ИС Information Resources on Inorganic Chemistry (IRIC), описывающей информационные ресурсы по неорганической химии и материаловедению. По своей сути IRIC является попыткой систематизации наиболее значимых ИС СНВМ [2]. Система реализована в виде веб-приложения и круглосуточно доступна по адресу http://iric.imet-db.ru/ на русском и английском языках.

Для решения второй задачи – обеспечения доступа к ИС СНВМ с возможностью быстрого поиска требуемой информации – необходима интеграция ИС в данной предметной области, что является не только большой организационной, но и технической проблемой.

3 Вербальное описание предметной области

Для успешной консолидации любых ИС необходимо, прежде всего, формализовать описание предметной области, которому должны соответствовать интегрируемые ИС.

Отличительной особенностью многих ИС СНВМ является узкая предметная направленность, обусловленная спецификой области исследования. Поэтому такие системы хранят информацию только о тех веществах и их характеристиках, которые относятся к исследуемой предметной области. В качестве примера можно привести ИС по фазовым диаграммам систем с полупроводниковыми фазами «Диаграмма» [3] и ИС по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами «Кристалл» [4]. Эти системы ориентированы на специалистов-материаловедов в области химии и материаловедения полупроводников и диэлектриков.

Таким образом, в разных информационных системах представлены различные характеристики (будем далее называть их свойствами) различных веществ и материалов (будем далее называть их сущностями). Значения свойств определяются, в первую очередь, составом неорганических веществ (набором химических элементов, входящим в их состав, и их соотношением, т. е. качественным и количественным составом), а также часто физические свойства зависят от кристаллической структуры образовавшейся твердой фазы. Поскольку ИС СНВМ тесно связаны с химией, то сущности в ИС СНВМ могут быть описаны с помощью иерархии понятий (система → вещество → модификация) в виде дерева (Рис. 1).

Обозначим сущности второго уровня общим термином «вещество», понимая под этим термином совокупность дискретных образований, обладающих массой покоя (т. е. атомы, молекулы и то, что из них построено). Итак, при описании химических сущностей можно использовать три уровня: система, вещество и кристаллическая (полиморфная) модификация (далее – модификация). При этом каждый последующий уровень *уточняет* (конкретизирует) информацию об описываемом химическом объекте.

Рисунок 1 Вершина иерархии понятий химических сущностей в неорганической химии

Приведем кратко определения основных терминов, использованных в иерархии понятий.

Химическая система (элементы, определяющие качественный состав) – система, образованная химическими элементами. Она может быть описана как множество атомов, образующих химическую систему. Более строго, химическая система – это совокупность микро- и макроколичеств веществ, способных под воздействием внешних факторов (условий) к превращениям с образованием новых химических соединений. Например, химическая система, в которую входят элементы медь, галлий и теллур, обозначается как Cu-Ga-Te.

Химическое соединение – однородное вещество постоянного или переменного состава с качественно отличным от свойств образующих его элементов химическим или кристаллохимическим строением. Соединение образовано из атомов нескольких химических элементов, связанных химической связью. На фазовой диаграмме область гомогенности соединения отделена (при всех температурах и давлениях) от области компонентов или твердых растворов на их основе. Элементы в соединении не могут быть разделены простым механическим способом, а лишь химической обработкой, нагреванием, электрическим током и т. д.

Раствор – макроскопически гомогенная смесь двух или более компонентов, состав которой при данных внешних условиях может непрерывно меняться в некоторых пределах.

Гетерогенная смесь – механическая смесь разнородных компонентов, в которой при заданных условиях отсутствует химическое взаимодействие.

Кристаллическая (полиморфная) модификация – форма пространственной организации твердого вещества.

Указанные выше химические определения являются в значительной степени нечеткими (размытыми). Поэтому иногда трудно провести

границу между, например, упорядоченным твердым раствором и соединением.

Необходимо отметить, что описание сущностей и их свойств в разных ИС по свойствам веществ происходит с разной степенью детализации. Так, например, в ИС «Диаграмма» описание большинства свойств химических сущностей ведется на уровне химических систем. А в ИС «Кристалл» некоторые свойства описаны на уровне химических веществ (например, температура плавления, растворимость и пр.), а некоторые свойства представлены на уровне конкретных модификаций (например, нелинейно-оптические коэффициенты, показатели преломления и пр.).

Очевидно, что свойства, указанные для химических сущностей на уровне систем, распространяются на все химические вещества этой системы и их модификации. Аналогично свойства, заданные на уровне химических веществ, распространяются на все химические модификации этого вещества. Данные замечания важны в контексте формального моделирования предметной области

4 Формальное описание предметной области в терминах теории множеств

При консолидации ИС возникают синтаксические и структурные конфликты из-за того, что ИС используют данные, различные по синтаксическому описанию и структуре. В ряде ИС используются реляционные системы управления базами данных (СУБД), в других – иерархические СУБД. В последнее время нередко строятся ИС, которые используют форматы JSON (JavaScript Object Notation), XML (eXtensible Markup Language) или какие-либо его известные приложения, например, RDF для хранения информации. В ИС, разработка которых велась довольно давно, нередко можно встретить собственные двоичные форматы для хранения и обработки данных. Все это многообразие моделей данных и схем представления, а также обработки информации приводит к тому, что ИС в том виде, в котором они существуют, зачастую являются несовместимыми с другими программными продуктами. Следует отметить, что изначально при проектировании ИС СНВМ взаимодействие с внешней программной средой не предусматривалось вовсе.

Разрешить синтаксические и структурные конфликты можно за счет введения общей схемы представления информации и обмена данными, построенной согласно описанию предметной области. Как уже было отмечено выше, при описании химических сущностей можно использовать три уровня: система, вещество и кристаллическая модификация. Указанная иерархия химических сущностей, которая рассматривается в контексте интегрированной ИС, представлена на Рис. 2.

Таким образом, в общую схему предметной области закладывается три типа объектов, соответствующих химическим сущностям: система (или химическая система – качественный состав вещества), вещество (количественный состав вещества) и модификация. При этом каждый последующий уровень *уточняет* (конкретизирует) описание объекта. Следовательно, все оболочки интегрируемых ИС СНВМ должны оперировать этими тремя типами объектов при ссылке на химические сущности. При этом стоит учитывать, что если характеризуется определенная кристаллическая модификация, то определена также и химическая система с веществом, модификация которого представляется, т.е. если описание химической сущности ведется на уровне модификаций, то все вышележащие уровни (вещество и система) считаются описанными. Следует заметить, что обратное неверно: при известном описании химической системы вещество и модификация не определены. Однако необходимо понимать, что при описании сущности на уровне системы все описанные свойства автоматически распространяются на все вещества и модификации, образованные в рамках этой системы. Это во многом напоминает наследование в объектно-ориентированном программировании (ООП).

Рисунок 2 Иерархия химических сущностей, рассматриваемая в контексте интегрированной ИС СНВМ

Воспользуемся теорией множеств для описания сущностей рассматриваемой предметной области, учитывая, что каждый последующий уровень в иерархии уточняет (дополняет) описание объекта. Обозначим множество химических систем S, множество химических веществ C, а множество кристаллических модификаций M. Тогда химическая система будет обозначаться s ($s \in S$), химическое вещество обозначим c ($c \in C$), а кристаллическую модификацию – m ($m \in M$).

Химическая система s может быть представлена как множество химических элементов e_i: $s=\{e_1,e_2,\ldots,e_n\}$. Химическое вещество c определяется не только множеством атомов (химических элементов), но и их количественным вхождением в состав вещества, раствора или смеси. Поэтому вещество c может быть представлено кортежем (s,f), где $s \in S$, а f является отображением множества атомов (химических элементов), которые образуют вещество, на множество пар $R^* \times R^*$, задающих соответственно минимальное и максимальное вхождения заданного элемента в вещество, раствор или смесь c. Значит, $f: e_i \rightarrow (R^*_{min}, R^*_{max})$, где $R^* = R^+ \cup \{x\}$.

R^+ – множество неотрицательных действительных чисел, а R^* – это множество R^+, расширенное элементом x. Элемент x служит для обозначения неизвестного числа, так как при обозначении смесей, где вхождение компонентов может варьироваться, принято использовать x для обозначения неизвестного, например, $Fe_{1-x}Se_x$. R^*_{min} и R^*_{max} – соответственно, минимальная и максимальная концентрации химического элемента e_i в веществе c. В случае, когда концентрация конкретного химического элемента e_i в веществе c фиксирована, $R^*_{min}=R^*_{max}$. Химическая модификация m может быть представлена кортежем (s, f, mod), где $s \in S$, $f: e_i \rightarrow (R^*_{min}, R^*_{max})$, а mod – строковое обозначение кристаллической модификации вещества, принятое в интегрированной ИС (одно из значений перечисления (enum) сингоний: {*Triclinic, Monoclinic, Orthorhombic, Tetragonal, Trigonal, Hexagonal, Cubic*}).

5 Формальное описание предметной области на объектно-ориентированном языке

При использовании объектно-ориентированного языка достаточно просто могут быть описаны формализмы предметной области, описанной выше. В качестве подтверждения данного тезиса рассмотрим формализацию с использованием языка C# (свободно доступная версия 6.0 https://github.com/vicdudarev/ChemicalHierarchy).

Не рассматривая детально предложенную реализацию, остановимся кратко на переходе от системы к веществу – дополнении информации о качественном составе количественным описанием. В предлагаемой реализации химическая система (класс ChemicalSystem) описывается в качестве одномерного массива типа ChemicalElement[], где ChemicalElement – класс для представления химического элемента (содержит обозначение элемента и его атомный номер). На уровне описания количественного состава вводится наследуемый от ChemicalSystem класс ChemicalSubstance, расширяющий описание количественным составом, представленным в виде одномерного массива типа Quantity[], где Quantity – простейший класс, содержащий пару значений Min и Max. Отметим, что в конструкторах классов выполняются все проверки на корректность задаваемых значений. Например, в конструкторе объектов класса ChemicalSubstance проверяется, что размер массива количественного описания совпадает с размером массива качественного описания, унаследованного от ChemicalSystem. Таким образом, развитые возможности объектно-ориентированных языков позволяют корректно реализовать предлагаемую в разделе 4 формализацию.

6 Представление свойств сущностей

Рассмотрев формализацию описания химических сущностей, перейдем к краткому изложению предлагаемого представления свойств химических сущностей. Как было отмечено, в интегрируемых ИС содержится информация по свойствам химических сущностей, например, плотность, растворимость, теплопроводность, ширина запрещенной зоны и т. п. При этом для каждой химической сущности в базе данных (БД) ИС нередко содержится несколько записей для описания значения свойства. Это обусловлено разными обстоятельствами. Во-первых, информация, содержащаяся в БД ИС, может быть взята из различных источников, при этом данные нередко расходятся. Это объясняется различными способами измерения, точностью измеряющей аппаратуры и т. д. Таким образом, в ИС СНВМ приводится несколько вариантов значения, например, плотности соединений. Во-вторых, значения рассматриваемых свойств зачастую зависят от внешних условий, при которых проводились измерения. Например, такие параметры, как растворимость и ширина запрещенной зоны, зависят от температуры. Другими словами, свойства часто являются функциями от различных аргументов, число которых, строго говоря, не фиксировано. Это означает, что разные свойства могут иметь разную структуру представления данных. Более того, одно и то же свойство в разных ИС СНВМ может фактически являться функцией от разного числа аргументов, и поэтому невозможно будет предложить универсальный формат представления заданного свойства для всех ИС. Это во многом может быть объяснено тем фактом, что при детальном исследовании какого-либо свойства число таких функциональных зависимостей от внешних параметров может возрастать. Следовательно, если такое свойство будет подробно рассмотрено в некоторой ИС СНВМ, которая еще не включена в общую интегрированную ИС, то при ее включении в состав интегрированной ИС возникнет проблема согласования форматов представления указанного свойства. Таким образом, невозможно заранее предусмотреть все зависимости и заложить их в общий формат представления данных для даже отдельно взятого конкретного свойства, не говоря о представлении свойств в целом.

В связи с вышеуказанным необходим некоторый механизм, позволяющий гибко представлять значения свойств в рамках интегрированной ИС. В настоящее время существует ряд широко используемых языков описания произвольных форматов данных, среди наиболее распространенных – JSON и XML. С помощью этих языков удобно описывать различные структуры данных, они являются межплатформенными форматами и поддерживаются большинством языков и библиотек [5]. На сегодняшний день представление данных с помощью таких языков является фундаментом для обеспечения взаимодействия различных программно-аппаратных платформ. В настоящее время все большее количество информации в современных промышленных системах представляется в форматах JSON и XML.

Использование этих форматов является целесообразным еще и потому, что они используются в качестве основы функционирования веб-сервисов.

Для разрешения семантических и структурных конфликтов необходимо стандартизировать форматы представления описанных химических сущностей и свойств в рамках интегрированной ИС на языках XML и JSON, т. е. необходимо разработать форматы соответствующих документов для представления химических сущностей, их свойств и другой информации. Это позволит обмениваться информацией между звеньями интегрированной ИС.

7 Заключение

Проблема интеграции ИС вообще и ИС СНВМ, в частности, чрезвычайно актуальна, поскольку доступ ко всей совокупности данных о веществах позволяет рассматривать такой консолидированный информационный источник в качестве объекта для всестороннего анализа и извлечения новых знаний.

В неорганическом материаловедении на первом этапе наиболее реалистичными являются попытки интеграции, основанные на учете специфики предметной области. Предложенное выше формальное описание предметной области – неорганического материаловедения – ни в коем случае не претендует на глубину проработки, которая бы удовлетворила материаловеда. В каждой из многочисленных областей материаловедения существует множество своих особенностей, учесть которые в большей или меньшей степени возможно при построении онтологий этих областей, основанных на сложных таксономиях.

Важно понимать, что сложность реализации ИС напрямую зависит от сложности формального описания предметной области. В этом смысле предложенная формальная модель (система → вещество → модификация), на наш взгляд, является приемлемым компромиссом между сложностью реализации интегрированной ИС и детальностью описания информации, представленной в отдельных интегрируемых ИС СНВМ.

Поддержка

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 16-07-01028, 17-07-01362 и 15-07-00980.

Литература

[1] Киселева, Н.Н.: Компьютерное конструирование неорганических соединений. Использование баз данных и методов искусственного интеллекта. М.: Наука (2005)

[2] Киселева, Н.Н., Дударев, В.А.: Информационная система по ресурсам неорганической химии и материаловедения. Вестник Казанского технологического университета, 17 (19), сс. 356-358 (2014)

[3] Христофоров, Ю.И., Хорбенко, В.В., Киселева, Н.Н. и др.: База данных по фазовым диаграммам полупроводниковых систем с доступом из Интернет. Изв. вузов. Материалы электронной техники, (4), сс. 50-55 (2001)

[4] Киселева, Н.Н., Прокошев, И.В., Дударев, В.А. и др.: Система баз данных по материалам для электроники в сети Интернет. Неорган. материалы, 42 (3), сс. 380-384 (2004)

[5] Christophides, I., Koffina, G., Serfiotis, V, Tannen, A.: Integrating XML Data Sources using RDF/S Schemas: The ICS-FORTH Semantic Web Integration Middleware (SWIM), Deutsch Dagstuhl Seminar: Semantic Interoperability and Integration (2004)