АНАЛИТИКА И УПРАВЛЕНИЕ ДАННЫМИ В ОБЛАСТЯХ С ИНТЕНСИВНЫМ ИСПОЛЬЗОВАНИЕМ ДАННЫХ

XVIII Международная конференция DAMDID / RCDL'2016

11–14 октября 2016 года, Ершово, Московская область, Россия

DATA ANALYTICS AND MANAGEMENT IN DATA INTENSIVE DOMAINS

XVIII International Conference DAMDID / RCDL'2016

October 11–14, 2016, Ershovo, Moscow Region, Russia

Федеральный исследовательский центр «Информатика и управление» РАН
Национальный исследовательский университет Высшая школа экономики
Национальный исследовательский ядерный университет «МИФИ»
Московская секция ACM SIGMOD
Российский фонд фундаментальных исследований

Аналитика и управление данными в областях с интенсивным использованием данных

XVIII Международная конференция DAMDID / RCDL'2016

Ершово, Московская обл., 11–14 октября 2016 года

Под редакцией
Л. А. Калиниченко, Я. Манолопулоса, С. О. Кузнецова

Federal Research Center "Computer Science and Control" of RAS
National Research University Higher School of Economics
Institute for Nuclear Power Engineering MEPhI
Moscow ACM SIGMOD Chapter
Russian Foundation for Basic Research

Data Analytics and Management in Data Intensive Domains

XVIII International Conference DAMDID / RCDL'2016

October 11–14, 2016, Ershovo, Moscow Region, Russia

Edit by
L. Kalinichenko, Y. Manolopoulos, S. Kuznetsov

ТОРУС ПРЕСС МОСКВА 2016

УДК [002:004.9] (063)
А 64
ББК [73+32.973.233]я431

А 64 Аналитика и управление данными в областях с интенсивным использованием данных: XVIII Международная конференция DAMDID / RCDL'2016 (11–14 октября 2016 года, Ершово, Московская обл., Россия): труды конференции / Под ред. Л. А. Калиниченко, Я. Манолопулоса, С. О. Кузнецова. – М.: ФИЦ ИУ РАН, 2016. 428 с.

ISBN 978-5-94588-206-5

Конференция «Аналитика и управление данными в областях с интенсивным использованием данных» ("Data Analytics and Management in Data Intensive Domains", DAMDID) представляет собой мультидисциплинарный форум исследователей и практиков из разнообразных областей деятельности людей, содействующий сотрудничеству и обмену идеями в сфере анализа и управления данными в областях исследований, движимых интенсивным использованием данных (ОИИД). Подходы к анализу данных и управлению данными, развиваемые в конкретных ОИИД X-информатики (таких как X = астро, био, гео, нейро, медицина, физика, химия, и пр.), социальных наук, а также различных ОИИД информатики, промышленности, новых технологий, финансов и бизнеса составляют предметную область конференции. Конференция DAMDID была образована в 2015 г. в результате трансформации конференции RCDL («Электронные библиотеки: перспективные методы и технологии, электронные коллекции», http://rcdl.ru) с сохранением преемственности по отношению к RCDL после многих лет ее успешного функционирования.

ББК [73+32.973.233]я431

Data Analytics and Management in Data Intensive Domains: 18th International Conference DAMDID / RCDL'2016 (October 11–14, 2016, Ershovo, Moscow Region, Russia): Conference Proceedings. Eds. L. Kalinichenko, Y. Manolopoulos, S. Kuznetsov. Moscow: FRC CSC RAS, 2016. 428 p.

ISBN 978-5-94588-206-5

The "Data Analytics and Management in Data Intensive Domains" conference (DAMDID) is planned traditionally as a multidisciplinary forum of researchers and practitioners from various domains of science and research, promoting cooperation and exchange of ideas in the area of data analysis and management in domains driven by data intensive research. Approaches to data analysis and management being developed in specific data intensive domains (DID) of X-informatics (such as X = astro, bio, chemo, geo, medicine, neuro, physics, etc.), social sciences, as well as in various branches of informatics, industry, new technologies, finance, and business constitute the universe of the conference discourse. DAMDID conference was arranged in 2015 as a result of transformation of the RCDL conference ("Digital Libraries: Advanced Methods and Technologies, Digital Collections" http://rcdl.ru) so that the continuity with RCDL has been preserved after many years of its successful work.

ISBN 978-5-94588-206-5

© Федеральный исследовательский центр «Информатика и управление» Российской академии наук, 2016
© ТОРУС ПРЕСС, 2016

Инфраструктура обеспечения данными специалистов в неорганической химии и материаловедении

© Н. Н. Киселева[1] © В. А. Дударев[1,2]

[1]Федеральное государственное бюджетное учреждение науки Институт металлургии и материаловедения им. А.А. Байкова Российской академии наук (ИМЕТ РАН),
[2]Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ), Москва

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

Аннотация

Проведен анализ реализуемых в мире крупных инфраструктурных проектов информационного обеспечения специалистов в области материаловедения (MGI, MDF, NoMaD и т.д.). Дан краткий обзор российских информационных ресурсов в области неорганической химии и материаловедения. Предложен проект инфраструктуры для обеспечения данными российских специалистов в этой области.

Авторы благодарят А.В. Столяренко, В.В. Рязанова, О.В. Сенько, А.А. Докукина за помощь в создании информационно-аналитической системы.

Работа выполнена при частичной финансовой поддержке РФФИ, проекты 16-07-01028, 14-07-00819 и 15-07-00980.

1 Введение

Конкурентные требования глобального рынка требуют постоянного обновления и улучшения потребительских свойств продукции. Качество и новизна выпускаемой продукции в значительной степени определяется материалами, используемыми при ее производстве. В связи с этим ускорение поиска, исследования и внедрения новых материалов с заданными функциональными свойствами является критически важной задачей развития промышленности и всей экономики стран в целом. В настоящее время, по мнению американских специалистов [1], между открытием нового материала и началом его практического использования проходит более 20 лет. Это связано с тем, что очень часто потребители не имеют достаточной информации даже об очень перспективных материалах, работы по исследованию и созданию технологии получения и обработки материалов необоснованно дублируются, используются не самые лучшие по потребительским и прочим параметрам вещества, что приводит к снижению качества продукции, росту затрат на ее производство и, в конечном счете, к утрате рыночной привлекательности выпускаемого продукта.

Одним из путей ускорения поиска, разработки и внедрения новых материалов является создание развитой инфраструктуры информационного обеспечения специалистов, в первую очередь, распределенной виртуально интегрированной сети баз данных и баз знаний, содержащих информацию о свойствах веществ и материалов и технологиях их получения и обработки, а также систем компьютерного конструирования и моделирования материалов, доступных из Интернет специалистам самого разного профиля: научным работникам, инженерам, технологам, бизнесменам, госслужащим, студентам и т.д.

В последние годы в развитых странах были выдвинуты и поддержаны правительствами инициативы, направленные на организацию инфраструктуры доступа к экспериментальным и расчетным данным о материалах. Краткий обзор некоторых инициатив ранее был дан в [2].

2 Стратегическая Инициатива Геном Материалов (Materials Genome Initiative (MGI))

В 2011 г в США была начата разработка проекта, названного Инициативой Геном Материалов (Materials Genome Initiative (MGI)) [3]. Цели MGI - ускоренное создание новых материалов, обладающих заданными свойствами, что критично для достижения высокого уровня конкурентоспособности промышленности США и будет способствовать поддержке их лидирующей роли во многих секторах современного материаловедения и промышленности: от энергетики до электроники, от обороны до здравоохранения. Особое внимание в MGI уделяется поддержке прорывных исследований в теории, моделировании свойств материалов и data mining как средств достижения существенного прогресса в материаловедении, что приведет к снижению затрат

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

на разработку, исследование и получение новых материалов. Задачи MGI – обеспечение разработки и внедрения новых материалов, в том числе и за счет координации исследований и предоставления доступа к расчетным моделям и инструментарию для оценки свойств и поведения материалов, а также использования прорывных методов моделирования и анализа данных. Реализация проекта MGI позволит создать механизмы, способствующие обмену данными и знаниями о материалах не только между исследователями, но и между академической наукой и промышленностью. Основой MGI является Инфраструктура инноваций в материаловедении (Materials Innovation Infrastructure), которая обеспечивает интеграцию методов и средств современного моделирования и экспериментальных исследований. Инфраструктура включает комплекс взаимосвязанных обслуживающих структур и объектов (в том числе и установок megascience), составляющих и/или обеспечивающих основу функционирования материаловедения как науки и прикладной области. На первом этапе на реализацию программы MGI выделено около 400 млн. долларов. В Подкомитет MGI Национального научно-технического Совета США (The National Science and Technology Council (NSTC)) входят представители Министерства обороны, Министерства энергетики, National Institute of Standards and Technology (NIST), National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), United States Geological Survey (USGS), Defense Advanced Research Projects Agency (DARPA) и т.д. [4]. Среди успешных поддержанных проектов этой инициативы можно выделить систему AFLOW [5, 6], содержащую БД с результатами квантовомеханических расчетов веществ и оснащенную компьютерным пакетом программ для проведения таких расчетов, и открытие нового вида высокопрочного и износостойкого стекла [7], осуществленного путем широкого использования теоретических расчетов.

3 Средства организации данных о материалах (The Materials Data Facility (MDF))

Учитывая важность материалов для достижения высокого уровня конкурентоспособности промышленности США, в июне 2014 г. национальный Консорциум сервисов данных (National Data Service (NDS)) объявил о пилотном проекте разработки средств для организации данных о материалах: The Materials Data Facility (MDF) [8], поддерживаемом NIST. Этот проект является ответом на инициативу MGI Белого дома по ускорению разработки современных материалов. MDF обеспечит материаловедов масштабируемым репозиторием для хранения экспериментальных и расчетных данных, в том числе и до их публикации, снабженных ссылками на соответствующие библиографические источники. MDF станет рычагом для создания национальной инфраструктуры коллективного использования информации, включая разработанные в мире БД по свойствам материалов и информационные системы для расчета и моделирования, а также будет способствовать организации обмена данными о материалах, в том числе и еще не опубликованными. Доступность данных и средств расчета обеспечивается современной информационной и телекоммуникационной инфраструктурой, которая позволяет предоставить данные исследователям материалов для многоцелевого использования, дополнительного анализа и проверки. Помимо NIST, среди исполнителей MDF необходимо выделить University of Chicago, Argonne National Laboratory, The University of Illinois, Northwestern University, Center for Hierarchical Materials Design и т.д. Репозиторий MDF сейчас включает [8], помимо многочисленных БД NIST [9], информационные системы с результатами квантовомеханических расчетов: AFLOW [5, 6], The Open Quantum Materials Database (OQMD) [10] и т.д.

4 Программа Поиска Новых Материалов (Novel Materials Discovery Laboratory (NoMaD))

Эта программа был ответом Евросоюза на американскую стратегическую инициативу MGI. Проект NoMaD [11, 12] направлен на создание Европейских центров превосходства (European Centres of Excellence) и предполагает разработку сети БД (Materials Encyclopedia) по свойствам веществ и материалов (в первую очередь, содержащих результаты расчетов), а также средств анализа этих данных и расчета веществ. Цель - ускорение разработки и использования материалов с заданными функциональными свойствами. Программа стартовала в ноябре 2015 г. в рамках проекта ЕС HORIZON2020 (объем финансирования около 5 млн. евро) [12]. Существенным недостатком NoMaD является ориентация на информационные ресурсы США (главным образом, БД NIST по свойствам веществ и материалов) и информационные системы с расчетными данными. В настоящее время репозиторий NoMaD [13] содержит только результаты квантовомеханических расчетов уже полученных соединений. Программа NoMaD во многом коррелирует с проектом Евросоюза Materials design at the eXascale (MaX) [14], включающим создание инфраструктуры для проведения квантовомеханических расчетов с использованием высокопроизводительных компьютерных систем (объем финансирования – свыше 4 млн. евро). Среди исполнителей NoMaD следует отметить ведущие организации Европы, такие как Humboldt University, Fritz-Haber-Institute of the Max Planck Society, King's College London, University of Barcelona, Aaalto University, Max Planck Institute for the Structure of Dynamics of Matter, Technical University of Denmark,

Max Planck Computing and Data Facility, Barcelona Supercomputing Centre и т.д.

5 Инициатива исследования материалов путем интеграции информации («Materials research by Information Integration Initiative» (MI2I))

Эта инициатива была предложена в 2015 г. японским правительством, которое создало на базе Национального института материаловедения (National Institute for Materials Science (NIMS)) Center for Materials Research by Information Integration [15]. В отличие от европейских программ созданный центр ставит своей задачей не только широкое использование квантовомеханических расчетов, но и поддержку и развитие имеющихся в Японии БД по свойствам веществ и материалов [16], их интеграцию с зарубежными информационными системами и применение методов искусственного интеллекта для прогноза новых веществ [17, 18].

6 Анализ реализуемых в мире крупных инфраструктурных проектов информационного обеспечения в области материаловедения

Следует отметить общие тенденции в разработке систем информационного обеспечения в материаловедческих областях:

- создание интегрированной сети БД по свойствам веществ и материалов;
- разработка и широкое применение расчетных методов;
- создание БД с расчетной информацией.

Анализ целей и предлагаемых в вышеуказанных инициативах методов и технологий их достижения показывает, что наиболее перспективны проекты США. Именно они позволят создать полноценную инфраструктуру информационной поддержки инновационной деятельности в разработке и внедрении новых материалов, обеспечив науку и промышленность достоверными и полными данными о свойствах веществ и материалов и разнообразным инструментарием (пакеты квантовомеханических расчетов, data mining и т.д.) для расчетов параметров веществ. Японская инициатива более ограничена. Она основана на использовании системы БД по свойствам веществ и материалов NIMS, а также использует имеющийся у исполнителей задел по применению уже известных расчетных методов (например, широко известного пакета квантовомеханических расчетов VASP [19]). Начаты работы по применению методов искусственного интеллекта [17, 18]. К тому же японские специалисты ограничили сферу деятельности материалами для электроники (источники питания, магнитные, термоэлектрические и спинтронные материалы) [15]. Проекты ЕС на их начальном этапе выглядят наименее перспективными. Ориентация на американские БД по свойствам веществ и только квантовомеханические расчеты значительно снижают потенциал и возможности этих инфраструктурных проектов. Тем не менее, следует отметить, что объективной предпосылкой для успешной реализации предложенных в США, ЕС и Японии инициатив являются, с одной стороны, успехи в разработке и применении методов расчета свойств веществ, и, с другой стороны, наличие множества баз данных по свойствам веществ и материалов, разработанных в последние годы в разных странах (обзор имеющихся БД в области неорганический химии и материаловедения дан в статье [20] и в БД IRIC (Information Resources on Inorganic Chemistry) [21]). Несмотря на то, что на создание и поддержку таких информационных систем затрачены сотни миллионов долларов, их использование экономически выгодно, т.к. они позволяют значительно сократить затраты на разработку новых материалов за счет уменьшения дублирования исследований и предоставления химикам и материаловедам оперативной и достоверной информации о свойствах веществ. В свою очередь, расчетные методы дают возможность еще до экспериментов оценить параметры веществ, указать перспективные для применений составы и разработать технологию получения и обработки материалов. Следствием решения этих задач является резкое сокращение затрат и времени на разработку и внедрение новых материалов. К сожалению, из-за введенных санкций против РФ, предполагаемой высокой цены доступа к разрабатываемым информационным системам, наличия в них информации о материалах и технологиях двойного назначения, доступ российских специалистов к этим информационным ресурсам будет крайне ограничен. Это может привести к серьезному отставанию в темпах разработки и внедрения новых материалов, что приведет к резкому уменьшению конкурентоспособности российской продукции, особенно в наукоемких отраслях. Единственный путь решения этой проблемы – это создание собственной инфраструктуры, обеспечивающей науку, образование, промышленность, бизнес, административные органы данными о материалах, технологиях их получения и обработки, сферах применения, производителях и потребителях материалов, а также средствами обработки и анализа накопленной информации, компьютерного моделирования и конструирования новых веществ и материалов, позволяющими принимать решения о выборе материалов для конкретных применений, перспективности разработки и использования конкретного вещества, о технологических особенностях производства, использования, утилизации и т.д. материалов и т.п. Средства, вложенные в такую импортозамещающую программу, достаточно быстро окупятся за счет сокращения затрат на разработку и исследование

новых материалов и прибыли от реализации наукоемкой продукции, конкурентной на мировом рынке.

7 Опыт разработки интегрированной информационной системы по свойствам неорганических веществ и материалов

Предпосылкой для успешного выполнения такого инфраструктурного проекта в России является опыт в разработке и интеграции БД по свойствам неорганических веществ и материалов, доступных из сети Интернет, также методов и программных средств для компьютерного конструирования новых веществ и материалов, основанных на использовании технологий data mining, и, в первую очередь, методов распознавания образов по прецедентам [20, 23]. Следует отметить, что интерес к применению методов data mining в неорганическом материаловедении связан с объективными трудностями, возникающими при квантовомеханических расчетах еще не полученных многокомпонентных неорганических веществ, особенно в твердой фазе. Например, для того, чтобы рассчитать электронную структуру неорганического соединения с использованием пакета VASP [19], необходимо знать его кристаллическую структуру, т.е. нужно получить и исследовать это соединение.

Рисунок 1 Схема информационно-аналитической системы для конструирования неорганических соединений

С помощью же методов распознавания образов, проанализировав имеющуюся информацию об уже известных веществах, хранящуюся в БД, можно прогнозировать еще не синтезированные соединения и оценивать некоторые их свойства, зная только хорошо известные параметры компонентов (химических элементов или более простых соединений). Для решения этой задачи в ИМЕТ РАН разработана специальная информационно-аналитическая система (ИАС) (рис.1), включающая интегрированную систему БД по свойствам неорганических веществ и материалов, подсистему поиска закономерностей в данных, прогнозирования новых соединений и оценки их свойств, базу знаний, базу прогнозов и другие подсистемы [22].

7.1 Интегрированная система баз данных по свойствам неорганических веществ и материалов

Интегрированная система баз данных по свойствам неорганических веществ и материалов в настоящее время объединяет информационные системы, разработанные в ИМЕТ РАН [20]: по фазовым диаграммам полупроводниковых систем («Диаграмма»), по свойствам акустооптических, электрооптических и нелинейнооптических веществ («Кристалл»), по ширине запрещенной зоны неорганических веществ («Bandgap»), по свойствам неорганических соединений («Фазы») и по свойствам химических элементов («Elements»), а также БД «AtomWork» по свойствам неорганических веществ, разработанную в National Institute for Materials Science (Япония), и БД ТКВ по термическим константам веществ, разработанную в ОИВТ РАН и МГУ.

БД по свойствам неорганических соединений «Фазы» [25, 26] в настоящее время содержит информацию о свойствах более 52 тыс. тройных соединений (т.е. соединений, образованных тремя химическими элементами) и более 31 тыс. четверных соединений, почерпнутую из более 32 тыс. литературных источников. Она включает краткую информацию о наиболее распространенных свойствах неорганических соединений: кристаллохимических (тип кристаллической структуры с указанием температуры и давления, выше которых реализуется указанная структура, сингония, пространственная группа, число формульных единиц в элементарной ячейке, параметры кристаллической решетки) и теплофизических (тип и температура плавления, температура распада соединения в твердой или газообразной фазах и температура кипения при атмосферном давлении). Помимо этого, БД содержит информацию о сверхпроводящих свойствах соединений. БД «Фазы» формируется на основе анализа сведений, почерпнутых из периодических изданий, справочников, монографий, отчетов, а также реферативных журналов (более половины источников хранятся в виде pdf-документов). Объем БД «Фазы» превышает 25 Гбайт, и она доступна

зарегистрированным пользователям из сети Интернет [25].

БД «Elements» [20, 26] включает информацию о 90 наиболее распространенных свойствах химических элементов: теплофизических (температура плавления и кипения при 1 атм, стандартные теплопроводность, молярная теплоемкость, энтальпия атомизации, энтропия и т.д.), размерных (ионные, ковалентные, металлические, псевдопотенциальные радиусы, объем атома и т.д.), других физических свойствах (магнитной восприимчивости, электропроводности, твердости, плотности и т.д.), положении в Периодической таблице элементов и т.д. БД доступна из сети Интернет [26].

БД «Диаграмма» [27, 28] содержит информацию, собранную и оцененную высококвалифицированными экспертами, о фазовых P-T-x-диаграммах двух- и трехкомпонентных полупроводниковых систем и о физико-химических свойствах образующихся в них фаз. Объем БД превышает 2 Гбайт. БД доступна зарегистрированным пользователям из сети Интернет [27].

БД «Bandgap» [29, 30] включает информацию (более 0.7 Гбайт) о ширине запрещенной зоны более 3 тыс. неорганических веществ. БД доступна пользователям из сети Интернет [30]. По предложению японской стороны эта БД будет интегрирована с японской БД «Computational Electronic Structure Database (CompES-X)», содержащей информацию об электронной структуре веществ [31].

БД «Кристалл» [32, 33] включает информацию о свойствах (пьезоэлектрических (пьезоэлектрические коэффициенты, упругие постоянные и т.д.), нелинейно-оптических (нелинейно-оптические коэффициенты, компоненты тензора Миллера и т.д.), кристаллохимических (тип кристаллической структуры, сингония, пространственная и точечная группа, число формульных единиц в элементарной ячейке, параметры кристаллической решетки), оптических (показатели преломления, область прозрачности и т.д.), теплофизических (температура плавления, теплоемкость, теплопроводность и т.д.) и т.д.), более 140 акустооптических, электрооптических и нелинейно-оптических веществ, собранную и оцененную высококвалифицированными экспертами в данной предметной области. Объем БД превышает 4 Гбайт. Она имеет русско- и англоязычную версии, доступные зарегистрированным пользователям из сети Интернет [32].

БД «Inorganic Material Database – AtomWork» [34, 35] содержит информацию о более чем 82 тыс. кристаллических структур, 55 тыс. значений свойств материалов и 15 тыс. фазовых диаграмм. БД доступна пользователям из сети Интернет [35].

БД по термическим константам веществ «ТКВ» [36] содержит доступную из сети Интернет информацию об около 27 тыс. веществ, образованных практически всеми химическими элементами.

7.2 Система компьютерного конструирования неорганических соединений

Основу системы компьютерного конструирования неорганических соединений составляют алгоритмы и программы распознавания образов по прецедентам, входящие в многофункциональную систему «РАСПОЗНАВАНИЕ», разработанную в ВЦ РАН [39] и объединяющую, помимо широко известных методов линейной машины, линейного дискриминанта Фишера, k-ближайших соседей, опорных векторов, нейросетевых и генетических алгоритмов, также уникальные алгоритмы, разработанные в ВЦ РАН: алгоритмы распознавания, основанные на вычисления оценок, алгоритмы голосования по тупиковым тестам, алгоритмы голосования по логическим закономерностям, алгоритмы статистического взвешенного голосования и т.д. В систему также интегрирована программа обучения ЭВМ процессу формирования понятий ConFor, разработанная в Институте кибернетики НАН Украины [38], в основу которой положена оригинальная организация данных в памяти ЭВМ в виде растущих пирамидальных сетей. Для отбора информативных свойств компонентов химических соединений в ИАС были включены программы, основанные на алгоритмах [39-41]. Использование методов распознавания образов позволило получить прогнозы тысяч новых неорганических соединений [22, 23, 29].

8 Проект инфраструктуры обеспечения данными российских специалистов в области неорганической химии и материаловедения

ИАС является, своего рода, пилотным проектом для создания информационной инфраструктуры для неорганического материаловедения. В ней виртуально интегрированы наиболее известные российские БД в этой области, а также начата их интеграция с зарубежными информационными системами. Большинство российских БД содержат ссылки на полные тексты публикаций, из которых извлечена информация, хранящаяся в БД. Подсистема компьютерного конструирования соединений позволяет найти закономерности в информации БД и использовать их для прогнозирования еще не полученных соединений и оценки их свойств. Следует отметить, что на этапе прогнозирования используются только данные о свойствах компонентов соединений (химических элементов или более простых соединений). Полученные прогнозы хранятся в специальной базе прогнозов, что расширяет функциональные возможности традиционных баз данных (пользователь получает не только известные

экспериментальные данные, но и прогнозы еще не синтезированных соединений и оценки некоторых их свойств).

При разработке российского проекта инфраструктуры информационного обеспечения специалистов в области неорганического материаловедения нужно учитывать все многообразие возможных запросов пользователей. Вполне естественно, что запросы академических ученых могут кардинально отличаться от запросов инженеров-конструкторов или производителей материалов. Однако общий проект информационной инфраструктуры должен включать в качестве необходимых элементов виртуально интегрированную систему российских и зарубежных баз данных по свойствам неорганических веществ и материалов, технологиям их получения и обработки, потребителям и производителям материалов и т.д., комплекс пакетов расчета и моделирования материалов, пользователями которых в большинстве случаев являются академические ученые, и виртуально интегрированную систему баз данных уже рассчитанных значений (рис. 2). Следует подчеркнуть, что технологии обработки, хранения, организации поиска необходимых сведений требуют разработки и использования самых современных программных средств и создания мощных центров обработки данных (ЦОД).

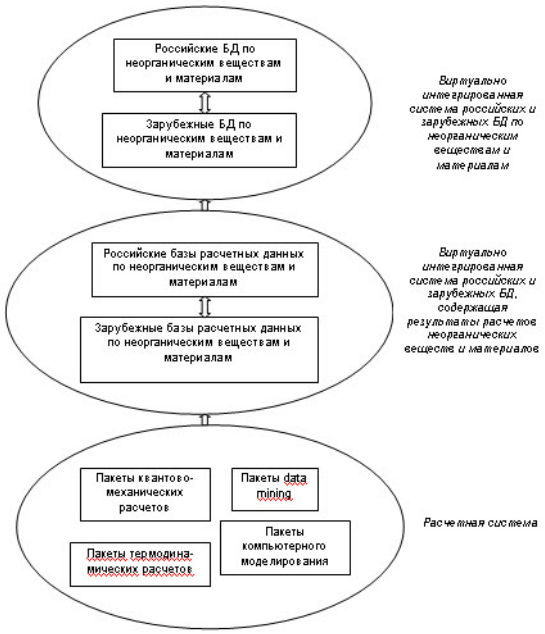

Рисунок 2 Схема инфраструктуры обеспечения данными российских специалистов в области неорганической химии и материаловедения

Система БД должна виртуально интегрировать наиболее важные для российских пользователей фактографические БД по неорганическим веществам и материалам (российские БД ИМЕТ РАН, ОИВТ РАН, МГУ и т.д и зарубежные БД NIMS [16], NIST [9], STN [42], Springer Materials [43], Materials Science International (MSI) [44] и т.д.), документальные БД ведущих издательских корпораций (Наука, Elsevier, Springer, Wiley, American Chemical Society, American Institute of Physics, Science и т.д.), а также базы еще не опубликованных в открытой печати сообщений (ВИНИТИ, ЦИТИС и т.д.), патентные базы (Роспатент, Questel, USPTO и т.д.), базы потребителей и производителей неорганических материалов и т.д. Необходимо выделять средства на ежегодное продление лицензий для пользования зарубежными базами и организовать единый портал бесплатного для российских пользователей доступа к ним (сейчас такие базы доступны для ограниченного числа организаций). Необходимо всячески поддерживать перевод в электронную форму коллекций наиболее известных в мире российских журналов, что несомненно будет способствовать повышению их авторитета и цитируемости.

Оснащение исследовательских организаций системами расчета необходимо начинать, в первую очередь, с обучения студентов и аспирантов пользованию наиболее известными пакетами квантовомеханических, термодинамических, статистических и т.д. расчетов. К сожалению, за последние четверть века отток специалистов в области теоретической химии сильно обескровил всемирно известные школы российских университетов и академических институтов, что значительно усложняет решение проблемы квалифицированного обучения молодых специалистов, а также использования расчетных методов. Нужно разрабатывать российские базы данных расчетных значений и интегрировать их с зарубежными информационными системами, которые сейчас еще открыто доступны в Интернет (например, [13, 30]), что позволит частично решить проблему квалифицированных расчетов веществ. Постановка экспериментов должна включать проведение или использование расчетов в качестве начального этапа исследований, что позволит сократить время и затраты на поиск и разработку новых материалов.

Важным компонентом разрабатываемого инфраструктурного проекта должна стать подсистема анализа запросов пользователей, особенно, специалистов в прикладных областях. Именно она позволит выявить группы материалов, на изучение которых нужно направить экспериментальные исследования. Статистика отказов в выдаче значений того или иного параметра вещества может стать стимулом к дополнительному экспериментальному исследованию запрашиваемого свойства.

9 Заключение

Переход российской экономики на инновационный путь развития и повышение конкурентоспособности продукции во многом определяется качеством, новизной и функциональными возможностями материалов. На

современном этапе развития технологий поиск, исследование и внедрение новых материалов требует создания развитой инфраструктуры, включающей академические организации с их потенциалом теоретических и экспериментальных исследований новых веществ, организации, ведущие прикладные исследования по разработке и внедрению новых материалов и технологий их получения и обработки, центры коллективного пользования с комплексами дорогостоящих установок, включая объекты megascience, и т.д. В последние годы в развитых странах инициированы стратегически важные для обеспечения технологического превосходства проекты (MGI, MDF, NoMaD и т.п.) создания инфраструктуры для ускорения разработки и внедрения новых материалов, обладающих заданными функциональными свойствами. Особое внимание в этих проектах уделено инфраструктуре информационного обеспечения. Российским ответом на стратегические инициативы США, ЕС, Японии в плане информационной инфраструктуры может являться создание федерального информационного центра, обеспечивающего специалистов информацией о свойствах веществ и материалов, технологиях их производства, а также расчетными данными, патентной информацией и т.д. В связи со спецификой предметной области основу такого информационного центра коллективного пользования должна составлять распределенная виртуально интегрированная сеть отечественных и зарубежных баз данных и баз знаний по веществам и материалам. Создание федерального информационного центра, интегрирующего информационные ресурсы в области материаловедения, будет способствовать резкому ускорению поиска, разработки и внедрения новых материалов, в сочетании со значительным сокращением затрат за счет уменьшения дублирования исследований и предоставления химикам и материаловедам оперативной и достоверной экспериментальной и расчетной информации о веществах и материалах.

Литература

[1] Materials Genome Initiative. Strategic Plan. National Science and Technology Council. Committee on Technology. Subcommittee on the Materials Genome Initiative. December 2014. 55 p. https://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/mgi_strategic_plan_-_dec_2014.pdf

[2] Калиниченко Л.А., Вольнова А.А., Гордов Е.П., Киселева Н.Н. и др., Проблемы доступа к данным в исследованиях с интенсивным использованием данных в России. Информатика и ее применения, 10(1), с. 3-23, 2016.

[3] Materials Genome Initiative for Global Competitiveness/ June 2011. 18 p. http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf

[4] Materials Genome Initiative. https://www.mgi.gov/partners

[5] Curtarolo S., Setyawan W., Wang S., et al. AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Comp. Mat. Sci., 58, p.227-235, 2012.

[6] Taylor R. H., Rose F., Toher C., et al. RESTful API for exchanging materials data in the AFLOWLIB.org consortium. Comp. Mat. Sci., 93, p. 178-192, 2014.

[7] Сайт University of Chicago http://www.uchicago.edu/features/microscopic_animals_inspire_innovative_glass_research/

[8] National Data Service. The Materials Data Facility. http://www.nationaldataservice.org/mdf/

[9] NIST Data Gateway. http://srdata.nist.gov/gateway/gateway?dblist=0

[10] Saal, J. E., Kirklin, S., Aykol, M., et al. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD), JOM. 65, p. 1501–1509, 2013.

[11] The Novel Materials Discovery (NOMAD) Laboratory. http://nomad-lab.eu/

[12] The Novel Materials Discovery (NOMAD) Laboratory. http://cordis.europa.eu/project/rcn/198339_en.html

[13] The NoMaD Repository. http://nomad-repository.eu/cms/

[14] Materials design at the eXascale. http://cordis.europa.eu/project/rcn/198340_en.html/

[15] Center for Materials Research by Information Integration. http://www.nims.go.jp/eng/research/MII-I/index.html

[16] NIMS Materials Database (MatNavi). http://mits.nims.go.jp/index_en.html

[17] Lee J., Seko A., Shitara K., Tanaka I. Prediction model of band-gap for AX binary compounds by combination of density functional theory calculations and machine learning techniques. Phys. Rev., B93(11), p. 115104, 2016.

[18] Toyoura K., Hirano D., Seko A., et al. Machine-learning-based selective sampling procedure for identifying the low-energy region in a potential energy surface: A case study on proton conduction in oxides. Phys. Rev., B93(5), p. 054112, 2016.

[19] VASP-site. https://www.vasp.at/

[20] Киселева Н.Н., Дударев В. А., Земсков В. С. Компьютерные информационные ресурсы неорганической химии и материаловедения. Успехи химии. 79(2), с. 162-188, 2010.

[21] БД IRIC (Information Resources on Inorganic Chemistry). http://iric.imet-db.ru/

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

[23] Киселева Н.Н., Дударев В.А., Столяренко А.В. Интегрированная система баз данных по свойствам неорганических веществ и материалов. Теплофизика высоких температур, 54(2), с. 228–236, 2016.

[24] Киселева Н., Мурат Д., Столяренко А. и др. База данных по свойствам тройных неорганических соединений «Фазы» в сети Интернет. Информационные ресурсы России, 4, с. 21-23, 2006.

[25] БД «Фазы». www.phases.imet-db.ru

[26] БД «Elements». http://phases.imet-db.ru/elements

[27] 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, с. 50-55, 2001.

[28] БД "Диаграмма". http://diag.imet-db.ru

[29] Киселева Н.Н., Дударев В.А., Коржуев М.А. База данных по ширине запрещенной зоны неорганических веществ и материалов. Материаловедение, 7, с. 3-8, 2015.

[30] БД «Bandgap». http://www.bg.imet-db.ru

[31] DB CompES-X. http://compes-x.nims.go.jp/index_en.html

[32] 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), с.380-384, 2004.

[33] БД "Кристалл". http://crystal.imet-db.ru

[34] Xu Y., Yamazaki M., Villars P. Inorganic Materials Database for Exploring the Nature of Material. Jap. J. Appl. Phys., 50(11), p.11RH02-1-5, 2011.

[35] БД "AtomWork". http://crystdb.nims.go.jp/index_en.html

[36] БД "ТКВ". http://www.chem.msu.su/cgi-bin/tkv.pl?show=welcome.html/welcome.html

[37] Журавлев Ю. И., Рязанов В. В., Сенько О. В. «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006.

[38] Гладун В. П. Процессы формирования новых знаний. София: СД "Педагог-6", 1995.

[39] Senko O. V. An Optimal Ensemble of Predictors in Convex Correcting Procedures // Pattern Recognition and Image Analysis, 19(3), p. 465-468, 2009.

[40] Yuan G.-X., Ho C.-H., Lin C.-J. An Improved GLMNET for L1-regularized Logistic Regression // J. Machine Learning Research, 13, p. 1999-2030, 2012.

[41] Yang Y., Zou H. A Coordinate Majorization Descent Algorithm for L1 Penalized Learning // J. Statistical Computation&Simulation, 84(1), p. 1-12, 2014.

[42] Сайт STN. http://www.stn-international.de/stn_home.html?&L=snavidlizydkyfaz

[43] Сайт Springer Materials. http://materials.springer.com/welcome

[44] Сайт MSI. http://www.msiport.com/.

Inorganic Chemistry and Materials Science Data Infrastructure for Specialists

Nadezhda N. Kiselyova, Victor A. Dudarev

World-wide materials science infrastructure projects are analyzed (MGI, MDF, NoMaD, etc.). Short overview of the Russian information resources on inorganic chemistry and materials science is given. Infrastructure project is proposed for Russian specialists in the domain to provide them with data.