Национальный исследовательский университет
«Высшая школа экономики»
Институт металлургии и материаловедения им. А. А. Байкова
Российской академии наук

В. А. Дударев

ИНТЕГРАЦИЯ ИНФОРМАЦИОННЫХ СИСТЕМ В ОБЛАСТИ НЕОРГАНИЧЕСКОЙ ХИМИИ И МАТЕРИАЛОВЕДЕНИЯ

МОСКВА

ББК 24.12 30.13 30.3 30–5–05 32.811 32.973–02

Настоящее издание осуществлено при финансовой поддержке Российского фонда фундаментальных исследований (проект № 16–17–00013)

Дударев Виктор Анатольевич

Интеграция информационных систем в области неорганической химии и материаловедения. — М.: КРАСАНД, 2016. — 320 с.

Книга посвящена вопросам создания интегрированных информационных систем в области неорганической химии. В работе приведен краткий обзор наиболее значимых фактографических баз данных по свойствам неорганических веществ, созданных в мире, и выполнен анализ архитектуры современных информационных систем в области материаловедения. Рассмотрены основные методы создания интегрированных систем (ETL, EII, EAI), и на их основе предложена комплексная методология интеграции материаловедческой информации, учитывающая требования пользователей и программ анализа данных. Приведены особенности реализации интегрированной информационной системы по свойствам неорганических веществ Института металлургии и материаловедения им. А. А. Байкова Российской академии наук (ИМЕТ РАН). В заключении рассмотрены примеры использования консолидированной информации для поиска закономерностей в данных и компьютерного конструирования новых перспективных неорганических соединений.

Книга предназначена для специалистов, занимающихся разработкой информационных систем по свойствам неорганических веществ. Изложенный материал позволяет грамотно реализовать структуры хранения информации, что открывает пути для дальнейшей интеграции с другими информационными системами и проведения интеллектуального анализа накопленных данных с целью прогнозирования свойств неорганических веществ.

ИЗДАНИЕ РФФИ НЕ ПОДЛЕЖИТ ПРОДАЖЕ

Издательство «КРАСАНД».
117335, Москва, Нахимовский пр-т, 56.
Формат 60×90/16. Тираж 300 экз. Печ. л. 20. Зак. №

Отпечатано в ОАО «Областная типография «Печатный двор».
432049, Ульяновск, ул. Пушкарева, д. 27.

ISBN 978–5–396–00745–1

17649 ID 217909

9 785396 007451

Оглавление

Введение

Необходимым условием инновационного развития промышленности является разработка и использование новых веществ и материалов. На текущем этапе развития материаловедение все чаще использует богатые информационно-прогнозирующие возможности, предоставляемые современными информационными технологиями. Для обеспечения химиков-технологов последними данными о свойствах и технологиях получения современных веществ создаются многочисленные специализированные информационные системы по свойствам неорганических веществ и материалов (ИС СНВМ). Разработка таких информационных систем ведется во всех промышленно развитых странах мира [1]. Одной из последних тенденций в данной области является организация круглосуточного удаленного доступа к ИС СНВМ с использованием телекоммуникационных сетей [2, 3]. Наиболее мощные ИС СНВМ, основанные на современных системах управления базами данных (СУБД), предлагают NIST (National Institute of Standards and Technology — Национальный институт стандартов и технологий, США), STN (The Scientific and Technical Information Network — Международная сеть научно-технической информации) и NIMS (National Institute of Materials Science — Национальный институт материаловедения, Япония) [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17]. Разработка ИС СНВМ в разных странах и организациях, как правило, происходит для решения узкого круга задач и без выработки единых стандартов представления информации, что значительно затрудняет попытки консолидации этих информационных систем. Дополнительным препятствием являются организационные трудности, т. к. большинство ИС СНВМ используются в коммерческих целях или являются открытыми для доступа пользователей только определенных стран или организаций.

Несмотря на существующие организационные трудности в последние годы наблюдается тенденция к кооперации в разработке ИС СНВМ и к интеграции уже созданных ИС, как на национальном, так и на международном уровне. Так, в рамках известной организации CODATA (http://www.codata.org) была создана специальная рабочая группа (Materials Task Group), занимающаяся развитием ИС СНВМ, которая объединяет крупных разработчиков информационных ресурсов в области материаловедения со всего мира. Одной из приоритетных задач данной рабочей группы является выработка стандартов для консолидации ИС СНВМ. Однако, несмотря

на предпринимаемые усилия, говорить об успехах в этой области преждевременно.

Актуальность решения задачи интеграции ИС вызвана стремлением устранить необоснованное дублирование работ по разработке и исследованию новых неорганических веществ и материалов. Кроме того, интеграция информации, содержащейся в ИС, по свойствам неорганических веществ и технологиям их получения, позволяет применять методы анализа для поиска взаимосвязей в данных. Использование найденных взаимосвязей позволяет проводить компьютерное конструирование новых перспективных соединений, обладающих заданными свойствами. Получаемая с помощью интегрированной ИС обобщенная информация может быть использована специалистами для поддержки принятия решений при выборе того или иного вещества и технологии его получения для использования в изделиях современной промышленности.

Автор выражает благодарность Н. Н. Киселевой за многолетнее плодотворное сотрудничество, а также В. Ф. Корнюшко, К. Ю. Колыбанову, Е. В. Бурляевой. Наконец, самых теплых слов заслуживает семья, которая всегда поддерживала и придавала сил и уверенности.

Исследования, представленные в настоящей монографии, выполнялись при частичной поддержке российских фондов и организаций: РФФИ (гранты № 04–07–90086, 06–07–89120, 05–03–39009, 12–07–09302, 09–01–12060, 09–07–00194, 12–07–00142 и 14–07–31032) и Правительства Москвы (гранты № 3–4 и 1.2.1 Программы «Инфраструктура и адресная поддержка науки»).

Глава 1
Особенности принятия решений при прогнозировании свойств неорганических веществ

Существующие определения термина «система поддержки принятия решений» (СППР), как правило, основаны на описании целей и функций этой системы [18]. Принятие решения в большинстве случаев заключается в генерации возможных альтернатив решений, их оценке и выборе лучшей альтернативы. Неопределенность является неотъемлемой частью процессов принятия решений. Одним из способов снятия неопределенностей является субъективная оценка специалиста (эксперта в предметной области), определяющая его предпочтения.

Лица, принимающие решения (ЛПР), вынуждены исходить из своих субъективных представлений об эффективности возможных альтернатив и важности различных критериев. Компьютерная поддержка процесса принятия решений основана на формализации методов получения исходных и промежуточных оценок, даваемых ЛПР, и алгоритмизации самого процесса выработки решения. Увеличение объема информации, поступающей ЛПР, усложнение решаемых задач, необходимость учета большого числа взаимосвязанных факторов и быстро меняющихся требований к решению настоятельно требуют использовать новый класс вычислительных систем — системы поддержки принятия решений (СППР).

Для разработки интегрированной информационной системы (ИС) для использования в СППР необходимо разработать структурную схему СППР. Эта разработка осуществлялась в соответствии с методикой проф. В. В. Кафарова [19] (рис. 1.1.1). Первым этапом этой методики является формулировка цели создания СППР. Основной целью создания СППР является обеспечение ЛПР прогнозами свойств соединений. Следующий этап — выделение подсистем СППР [20]. Вначале сложный программный комплекс СППР разбивается на ряд подсистем. Следующим шагом является выделение информационных связей. И на последнем этапе определяются управляющие воздействия ЛПР (рис. 1.1.1). Следует отметить, что на основе полученных прогнозов ЛПР принимает решение

о проведении экспериментальной проверки свойств соединений, прогнозируемых СППР. Более подробно процесс принятия решений рассматриваются в седьмой главе настоящей работы.

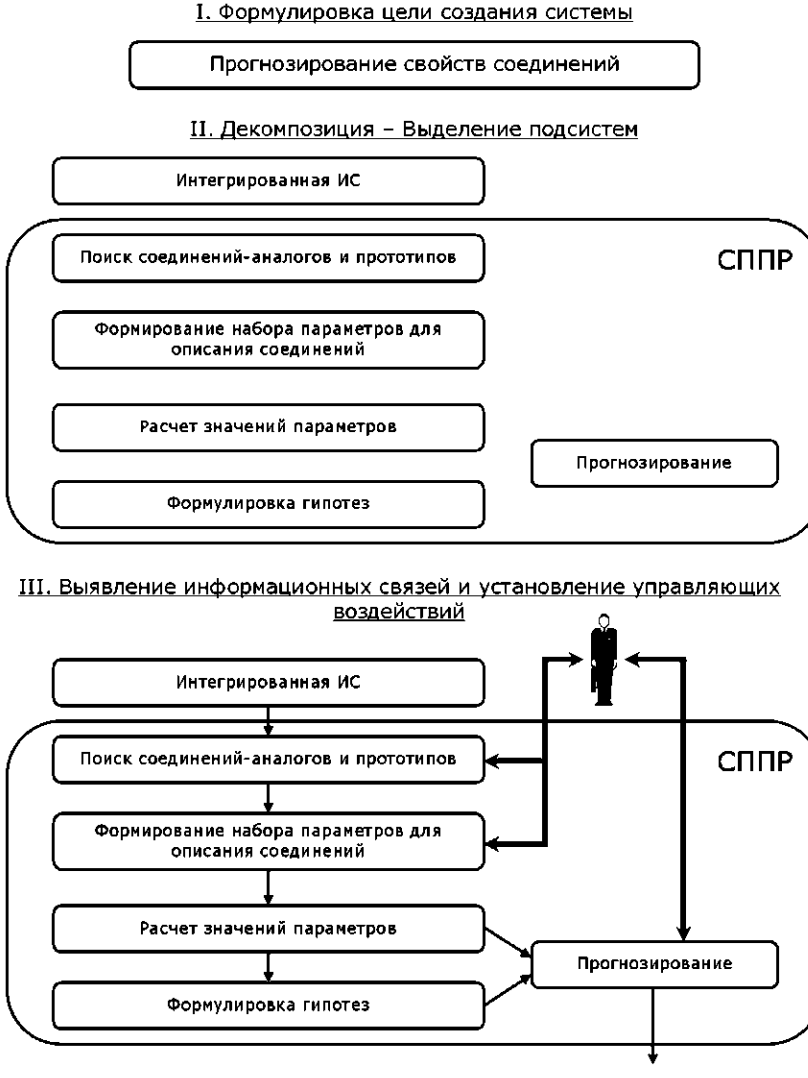


Рис. 1.1.1. Системный подход к разработке структурной схемы СППР

Целью настоящей работы является решение проблемы информационной поддержки компьютерного конструирования неорганических соединений на основе интеграции гетерогенных информационных систем по свойствам веществ и технологиям их получения. Для ее достижения необходимо выполнить обзор ИС по свойствам веществ и технологий их получения для промышленности, современных методов интеграции ИС, способов конструирования неорганических соединений.

1.1. Способы конструирования неорганических соединений

На современном этапе развития вычислительных систем наблюдается их повсеместное использование для обработки больших массивов данных и осуществления ресурсоемких вычислений. Не исключением является и современное материаловедение, где к помощи компьютеров прибегают не только для моделирования различных физико-химических процессов, но и для осуществления прогнозирования. Термин «компьютерное конструирование» (computer-assisted design) впервые появился в семидесятых годах прошлого века в работах Кори и Уипке применительно к синтезу сложных органических соединений с помощью компьютера [21]. Соблюдение правил валентности для углерода и водорода упрощает решение задачи компьютерного конструирования органических соединений, в отличие от неорганических соединений, где правило валентности соблюдается не для всех видов химических связей между атомами. Термин «компьютерное конструирование неорганических соединений», появившийся в девяностые годы прошлого века, обозначал поиск качественного и количественного состава соединений, которые еще не были синтезированы, а также оценку их свойств. В нашей стране подобными работами занимается научная группа под руководством Н. Н. Киселевой (ИМЕТ РАН) с использованием современных программных комплексов [1], разработанных с привлечением специалистов ВЦ РАН и Института кибернетики НАН Украины. По ее определению, «компьютерное конструирование неорганических соединений» заключается в нахождении совокупности химических элементов и их соотношения для создания определенной молекулярной или кристаллической пространственной структуры соединения, позволяющей реализовать необходимые функциональные свойства.

С помощью методов компьютерного конструирования неорганических соединений на текущем этапе решаются следующие типы задач:

- образование (отсутствие образования) соединений в химической системе;
- образование (отсутствие образования) соединений заданного количественного состава в химической системе;
- прогнозирование типа кристаллической структуры;
- интервальное прогнозирование значений свойств неорганических соединений.

Для решения указанных задач известны следующие подходы:

- квантовомеханический подход, основанный на решении уравнения Шрёдингера или его обобщений (уравнение Клейна—Гордона, уравнение Паули, уравнение Дирака и т. п.);
- простейшие эмпирические двух- и трехмерные критерии образования соединений с заданными свойствами, (например, фактор толерантности Гольдшмидта, правило Лавеса);
- многомерные эмпирические классифицирующие закономерности, получаемые с помощью методов компьютерного распознавания образов в N-мерном пространстве признаков.

1.1.1. Методы квантовой механики

Использование квантовой механики позволяет в теории рассчитать свойства любого химического соединения. Для этого требуется решить основное уравнение — уравнение Шрёдингера:

$$\hat{\mathbf{H}}\Psi_n = E_n\Psi_n,$$

где Ψ_n — собственная функция, содержащая в себе всю информацию о свойствах системы, $\hat{H} = T + V$ — гамильтониан, определяющий полную энергию системы, равную сумме оператора кинетической энергии T всех частиц системы и оператора их потенциальной энергии V. E_n — полная энергия системы [22].

Отмечается, что точное решение уравнения Шрёдингера возможно только для атома водорода и гипотетического иона гелия He^+. С использованием численных методов можно получить значения E_n и Ψ_n с любой заранее заданной точностью. Однако такое решение становится не только экономически неприемлемым из-за огромных затрат на расчеты, но и практически невозможным [23].

В связи с невозможностью точного численного решения уравнения Шрёдингера стали появляться приближенные (полуэмпирические) методы квантовой химии. В данных методах большую роль играет правильный выбор приближения для каждого конкретного случая и интерпретация

полученных результатов. Все приближенные методы решения уравнения Шрёдингера можно разделить на три основные группы [24]:

- адиабатическое приближение (метод Борна—Оппенгеймера), при котором движения ядер отделяются от движения электронов;
- одноэлектронное приближение, заменяющее локальное взаимодействие между электронами некоторым средним взаимодействием;
- линейная комбинация атомных орбиталей (МО ЛКАО), при которой электронная функция многих центров заменяется конечной суммой одноцентровых функций.

По результатам анализа многочисленных квантовомеханических расчетов делается вывод о неприменимости этого подхода для расчета характеристик еще не полученных соединений. Проблемы расчета параметров новых неорганических соединений не могут быть сведены только к математическим сложностям приближенного численного решения уравнения Шрёдингера, т. к. трудности расчета новых неорганических соединений являются следствием природы этих материальных объектов. На основании этого делается вывод [1] о текущей неприменимости методов квантовой механики для прогнозирования образования новых соединений и предсказания их свойств.

1.1.2. Простейшие эмпирические зависимости

Многочисленные сложности, возникающие при попытках квантовомеханических расчетов сложных химических соединений исходя из свойств их элементов, привели к появлению эмпирических критериев для классификации известных веществ. На базе этих критериев в дальнейшем проводились попытки экстраполяции найденных зависимостей для прогнозирования свойств неизученных объектов. Так был совершен переход к априорному прогнозированию новых соединений или их свойств с использованием полученных ранее эмпирических критериев.

В качестве примеров известных эмпирических критериев можно привести критерий Маттиаса для прогноза новых сверхпроводников с кристаллической структурой типа A15, правила Юм—Розери для определения способности химического элемента растворяться в металле с образованием твердого раствора; диаграммы Даркена—Гурри для прогноза взаимной растворимости металлов; правило Лавеса для предсказания кристаллической структуры некоторых интерметаллических соединений [25]. Следует подчеркнуть, что разработанные критерии являются результатом анализа огромного массива экспериментальных данных, а не результатом каких-либо теоретических расчетов. Более того, в большинстве случаев теоретическая физика не способна объяснить причину

успешного выполнения тех или иных критериев, носящих эмпирических характер.

С точки зрения конструирования новых неорганических соединений, основной способ разработки эмпирических критериев должен заключаться в подборе таких свойств химических элементов или функций от этих свойств, которые бы образовывали пространство (желательно, с минимальным числом измерений), в котором известные вещества образовывали непересекающиеся области (кластеры). Основным достоинством таких критериев является их простота, позволяющая построить наглядные проекции в полученном пространстве свойств, например, в виде двух или трехмерных проекций. Но не следует забывать и о существенных недостатках:

- высокая трудоемкость разработки эмпирических критериев;
- эмпирические критерии с малой размерностью признакового пространства не могут учитывать всю совокупную сложность химических соединений, принадлежащих к разным классам веществ;
- эмпирические критерии могут утратить прогнозирующую способность в результате появления новых данных, которые не согласуются со старым опытом (заметим, что изменение критериев часто либо невозможно, либо связано с большими затратами).

Стоит отметить, что естественным развитием эмпирических критериев с малой размерностью являются сложные многомерные критерии, поиск которых до эры бурного развития вычислительных систем был не возможен.

1.1.3. Многомерные классифицирующие правила

Переход от поиска простых эмпирических критериев к более сложным оказался возможным только с использованием современных вычислительных систем, оснащенных специальными программами анализа больших массивов данных. Поиск многомерных закономерностей является результатом эволюции рассмотренных ранее эмпирических методов. Применение компьютеров и программ поиска многомерных классифицирующих закономерностей в больших объемах экспериментальной информации позволяет резко сократить время разработки новых критериев и видоизменения старых критериев в связи с появлением новых данных, вступающих в противоречие с найденными взаимосвязями. При этом размерность критериев ограничивается только вычислительной мощностью современных компьютеров и возможностями программ анализа больших массивов данных.

С точки зрения простоты преимущества одно- и двухмерных критериев становятся несущественными после появления ЭВМ, с помощью

которых можно относительно быстро спрогнозировать новые вещества, используя многомерные классифицирующие правила. Средства современной визуализации любой проекции многомерного пространства признаков позволяют исследователям анализировать полученные результаты.

Появление концепции «черного ящика» позволило подойти к решению сложно формализуемых задач, в которых исследователь располагает только набором входных и выходных параметров, но не знает, каким образом входные параметры влияют на результат. Совокупность подобных задач и методов их решения называют *анализом данных* (*data analysis*, или *data mining* [26]). К недостаткам этих методов, с точки зрения любой предметной области, можно отнести не только недостаточную строгость полученных моделей, но и частую невозможность их интерпретации, например в случае использования самообучающихся нейронных сетей. Однако при отсутствии хорошо работающих теоретических методов (см. квантовомеханический подход), эти методы являются, пожалуй, единственно возможным вариантом получить некоторую модель происходящих процессов. Другая противоположность — полный отказ от таких «нестрогих» методов и использование только экспериментов в неорганической химии — окажется слишком затратным.

По сути, поиск многокритериальных классифицирующих правил возможен только при использовании больших массивов фактографических данных по свойствам веществ и материалов. Наличие такого массива данных означает автоматическое использование БД по свойствам неорганических веществ и материалов. Т. е. приходим к использованию информации из материаловедческих БД для поиска взаимосвязей. Данный подход широко известен — Knowledge Discovery in Databases. Сам подход не задает набор методов обработки или пригодные для анализа алгоритмы, он определяет последовательность действий, которую необходимо выполнить для того, чтобы из исходных данных получить знания, пригодные для дальнейшего использования. Данный подход универсальный и не зависит от предметной области, что является его несомненным достоинством.

Knowledge Discovery in Databases (KDD) — это процесс поиска полезных знаний в «сырых данных». KDD включает в себя вопросы: подготовки данных, выбора информативных признаков, очистки данных, применения методов data mining, постобработки данных и интерпретации полученных результатов. Безусловно, основным звеном всего этого процесса являются методы data mining, позволяющие обнаруживать сложные взаимосвязи в данных.

Процесс Knowledge Discovery in Databases может быть представлен в виде набора следующих шагов (рис. 1.1.2):

- **Подготовка исходного набора данных**. Этот этап заключается в создании набора данных, в том числе из различных источников, выбора

обучающей выборки и т. д. Для этого должны существовать развитые инструменты доступа к различным источникам данных.

- **Предобработка данных**. Для эффективного применения методов data mining следует обратить внимание на вопросы предобработки данных. Данные могут быть избыточны, недостаточны и т. д. Данные могут содержать пропуски, шумы, выбросы и т. д. Данные должны быть качественны и корректны с точки зрения используемого метода data mining. Поэтому второй этап KDD заключается в предобработке данных. Если размерность исходного пространства очень большая, то желательно применять специальные алгоритмы понижения размерности. Под последним понимается как отбор наиболее информативных признаков, так и отображение данных в пространство меньшей размерности.

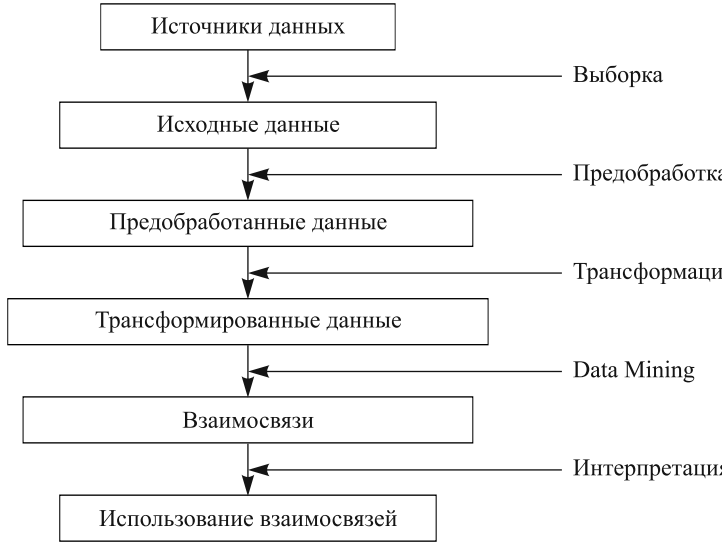


Рис. 1.1.2. Процесс Knowledge Discovery in Databases

- **Трансформация и нормирование данных**. Этот шаг необходим для приведения информации к пригодному для последующего анализа виду. Необходимо проделать такие операции, как приведение типов, квантование, нормирование и прочее. Кроме того, некоторые методы анализа требуют, чтобы исходные данные были представлены в некотором определенном виде. Например, нейронные сети работают только с числовыми данными, причем они должны быть нормированы.

- **Нахождение закономерностей (data mining)**. На этом шаге применяются различные алгоритмы для нахождения сложных взаимосвязей в данных. Информация, найденная в процессе применения методов data mining, должна быть нетривиальной и ранее неизвестной. Найденные взаимосвязи должны описывать связи между свойствами, предсказывать значения одних признаков на основе других и т. д. В этом смысле найденные взаимосвязи можно рассматривать как новые знания, которые должны быть применимы и для новых данных с некоторой степенью достоверности.
- **Постобработка данных**. Интерпретация результатов и применение полученных знаний.

При использовании KDD наиболее важным является этап подготовки данных и выбора алгоритмов для поиска взаимосвязей в данных.

1.2. Математические методы распознавания

Для поиска взаимосвязей в больших массивах данных часто применяют математические методы интеллектуального анализа данных, известные также как распознавание образов. Термин «распознавания образов» обязан своим появлением американскому ученому Фрэнку Розенблатту, который в 1960 году создал устройство, реализующее физиологическую модель зрения [27]. Свою распознающую машину он назвал персептроном (от латинского *percepto* — понимаю, познаю). Персептрон распознавал (различал, опознавал) зрительные образы. Так появился термин — *распознавание образов*. При развитии данного направления решались не только задачи распознавания изображений, но и другие задачи, которые было сложно формализовать, используя математические модели. Распознавание образов использовалось при обработке зашумленных сигналов, речи и др. Основная черта входных данных заключалась в их неполноте, слабой структурированности и противоречивости, т. е. во всем, что затрудняет попытки формализации.

За полвека, прошедшие со времени создания персептрона, интеллектуальный анализ данных и распознавание образов сильно развились и нашли широкое применение. В обзоре затруднительно очертить все проблемы интеллектуального анализа данных и области использования его математических методов в различных сферах деятельности человека. Несмотря на то, что большинство применений методов распознавания образов относится не к распознаванию изображений, а к решению задач классификации, прогнозирования, идентификации, до сих пор ученые используют термины, пришедшие в эту область из работ основоположников.

На текущем этапе развития область, именуемая интеллектуальным анализом данных, решает следующий ряд задач:

- кластер-анализ (автоматическая классификация или распознавание образов без учителя);
- поиск наиболее важных классифицирующих признаков (параметров классификации);
- распознавание образов и прогнозирование (классификация с учителем);
- поиск данных, существенно отклоняющихся от выявленных взаимосвязей (анализ аномалий);
- построение коллективных решений в задачах классификации (комитетные методы).

В настоящей работе системы интеллектуального анализа данных используются для автоматического поиска нелинейных зависимостей в данных. Сейчас технологии анализа данных бурно развиваются — не только создаются программные продукты для анализа данных, но и сами средства анализа данных встраиваются во все новые продукты. Например, вместе с СУБД Microsoft SQL Server поставляется Business Intelligence Development Studio. Но даже самые лучшие программные средства никогда не заменят специалиста-химика, способного провести всесторонний анализ наблюдаемых явлений. Таким образом, современные компьютерные технологии являются хорошим помощником химика, осуществляющего поиск сложных взаимосвязей в данных с целью поиска новых закономерностей, в значительной мере упрощая его работу.

1.2.1. Формальная постановка задачи прогнозирования

Прежде чем приступить к формальной постановке задачи прогнозирования свойств неорганических веществ, приведем ряд соглашений. Под распознаванием образов будем понимать задачу классификации объектов по нескольким классам с учителем, то есть классификация основывается на прецедентах. Таким образом, **прецедент** — это объект, принадлежащий обучающей выборке, правильная классификация которого известна. То есть прецедент принимается за образец при решении задачи классификации. Стоит заметить, что идея принятия решений на основе прецедентности — основополагающая в естественнонаучном мировоззрении. Считается, что все объекты или явления разбиты на конечное число классов, а для каждого класса, в свою очередь, известно конечное число объектов-прецедентов. Измерения, или свойства, используемые при классификации объектов, называются **признаками**. Таким образом, признак — это некоторое коли-

чественное или качественное свойство объекта произвольной природы. Совокупность признаков, которые относятся к одному объекту, называется **вектором признаков** или **признаковым описанием** объекта. Вектора признаков принимают значения в пространстве признаков. В рамках задачи распознавания образов считается, что каждому объекту ставится в соответствие один и только один вектор признаков. И наоборот: каждому значению вектора признаков соответствует только один объект. **Классификатором** или решающим правилом называют правило отнесения объекта к одному из известных классов на основании его вектора признаков. Рассмотрим формальную постановку задачи классификации (обучение с учителем).

Описанием объекта является **вектор признаков:** $x \in X \subseteq R^n$**.**

Классом называется некоторое подмножество объектов

$$K_y = \{x \in X \mid y^*(x) = y\}$$

множества X. Пусть $y \in Y \subseteq Z$ — множество маркеров классов.

Функция $X \xrightarrow{y^*} Y$ — отображение, определенное для всех $x \in X$, задающее разбиение X на подмножества K_y.

Обучающей выборкой называется набор пар (набор прецедентов) $S = (x_1, y_1), \ldots, (x_l, y_l)$, для которых $y^*(x_i) = y_i$, $i = 1, \ldots, l$. (l — размер обучающей выборки), то есть это известная информация об отображении $X \xrightarrow{y^*} Y$.

Основной гипотезой для применения алгоритмов классификации в распознавании образов является предположение, что множество $X \times Y$ является вероятностным пространством с вероятностной мерой P. Прецеденты $(x_1, y_1), \ldots, (x_l, y_l)$ появляются случайно и независимо в соответствии с распределением P.

Таким образом, задача классификации заключается в построении функции-классификатора $F(x)$, приближающей отображение y^*, основываясь на обучающей выборке $(x_1, y_1), \ldots, (x_l, y_l)$.

Несколько дополнительных определений:

Эмпирическим риском называется $P(F(x) \neq y \mid (x, y) \in S)$, то есть вероятность неверной работы классификатора для объектов из обучающей выборки.

Общим риском называется $P(F(x) \neq y \mid x \in X)$, то есть вероятность того, что классификатор ошибется на данных, не входивших в обучающую выборку. Основной целью при построении функции-классификатора является минимизация общего риска. Поскольку напрямую вычислить величину общего риска невозможно, для проверки качества классификатора используется оценка ошибки на **контрольной выборке**, которая состоит из прецедентов, не входящих в обучающую выборку. Говорят, что классифи-

катор обладает хорошей **обобщающей способностью,** если при обучении классификатор эффективно уменьшает общий риск, оцененный на контрольной выборке. Таким образом, обучение можно рассматривать как процесс построения классификатора по обучающей выборке *S*.

Говорят, что алгоритм построения склонен к **переобучению**, если при минимизации эмпирического риска, общий риск начинает возрастать. Это связано с тем, что классификатор начинает обобщать признаки, свойственные не данным в целом, а конкретным прецедентам из обучающей выборки. Переобучению способствует невыполнение принятой гипотезы, шум, высокая сложность классифицирующей функции (высокая размерность Вапника—Червоненкиса [28]).

Применительно к компьютерному конструированию неорганических соединений задача распознавания может быть дана следующим образом. Пусть каждое неорганическое соединение описано вектором признаков, имеющим составную структуру

$$x = (x_1^{(1)}, x_2^{(1)}, ..., x_M^{(1)}, x_1^{(2)}, x_2^{(2)}, ..., x_M^{(2)}, ..., x_1^{(L)}, x_2^{(L)}, ..., x_M^{(L)}),$$

где L — количество химических элементов в соединении, а M — количество параметров химических элементов, выбранных для описания ($x \in X \subseteq R^n$). Каждое соединение также характеризуется принадлежностью к определенному классу: $K_y = \{x \in X \mid y^*(x) = y\}$. Обучающая выборка S состоит из l объектов: $S = \{x_i, i = 1, ..., l\}$. По сути, обучающая выборка S является матрицей, содержащей экспериментальные данные, причем каждая строка этой матрицы соответствует описанию химического объекта (является прецедентом). Примерами объектов могут быть оксидные системы, химические соединения и т. д.

Образ — множество (класс) любых объектов, объединенных общими свойствами. Например, образом могут быть: класс тройных систем с образованием соединений определенного состава или класс соединений с кристаллической структурой типа шпинели и т. п. Цель обучения — построить классифицирующие правила, которые позволяют не только отличить химические объекты разных классов, но и обладают прогностической способностью образовывать новые комбинации химических элементов, которые не использовались для обучения, и относить их к одному из K_y классов. Как видно, особенность предметной области проявляется в формировании признакового описания, имеющего составную структуру: набор свойств химических элементов (компонентов неорганического вещества) повторяется L раз.

Стоит отметить, что неорганическое материаловедение как предметная область имеет свою специфику, которая создает дополнительные трудности при использовании математических методов распознавания [1]. Основными проблемами являются следующие:

- Малая информативность признаков — свойств химических элементов.
- Сильная закоррелированность признаков. Как следствие периодического закона — все свойства элементов находятся в периодической зависимости от общего параметра — атомного номера химического элемента.
- Наличие пропусков в признаках (отсутствующие значения в данных), которые могут быть вызваны как слабой изученностью свойств, так и принципиальным отсутствием некоторых значений. Например, отсутствие температуры плавления для углерода при атмосферном давлении, т. к. при нагревании при нормальном давлении углерод возгоняется.
- Частая асимметрия в размерах классов обучающей выборки.
- Возможность экспериментальных ошибок в обучающих выборках, как при классификации, так и при указании значений признаков.

В процессе тестирования различных алгоритмов обучения ЭВМ на конкретных задачах были выработаны критерии отбора программ для химических приложений. Программа анализа данных должна обеспечивать:

- возможность успешного анализа, как больших выборок, так и возможность определения качественных классифицирующих закономерностей при малых обучающих выборках;
- возможность работы в условиях слабого выполнения гипотезы компактности (т. е. при «размытости» и взаимном перекрытии границ классов в признаковом пространстве);
- возможность работы с качественными свойствами и пропусками в некоторых значениях свойств;
- быстрое обучение и прогнозирование.

1.2.2. Методы обучения ЭВМ распознаванию образов

Рассмотрим некоторые из наиболее широко используемых методов распознавания образов [184], которые используются специалистами для компьютерного конструирования неорганических веществ.

1.2.2.1. Методы прикладной статистики

Традиционные методы прикладной статистики решают три основных проблемы:

- Статистическое исследование структуры и характера взаимосвязей, существующих между анализируемыми количественными переменными. К нему относятся: корреляционный, факторный, регрессионный анализ, анализ временных рядов [29].

- Методы классификации объектов и признаков, например дискриминантный и кластерный анализ [30].
- Снижение размерности исследуемого признакового пространства с целью более лаконичного объяснения природы анализируемых данных, например метод главных компонент [31] и многомерное шкалирование [32].

Благодаря многообразию существующих статистических методов список решаемых ими проблем и сфер их применения является поистине неисчерпаемым. Статистические модели требуют наличия полной априорной информации, на основе которой могут быть определены вероятностные характеристики классов, что весьма затруднительно при решении определенных химических задач, явно носящих прецедентный характер. Данные задачи отличаются тем, что априорная информация о представительности некоторых классов отсутствует. Так, в ряде случаев классы могут быть представлены в виде единичных прецедентов (в одном классе один или два объекта). К недостаткам статистических методов нередко относят повышенные требования к математической подготовке пользователя.

1.2.2.2. Методы рассуждения по аналогии

Основной принцип, заложенный в методы рассуждения на основе аналогичных случаев (CBR — case based reasoning), является достаточно простым: чтобы осуществить прогнозирование или выбор правильного решения, проводятся поиск аналогичных (или близких) прецедентов имеющейся ситуации и выбирается тот же ответ, который был правильным для этих прецедентов.

Метод k ближайших соседей

Данный метод (k-nearest neighbors, k-NN) в теории распознавания, без сомнений, является классическим. Параметром метода является число k — обозначающее число «соседей» у распознаваемого объекта u. Распознаваемый объект относится в тот класс, из которого он имеет максимальное число «соседей». Оптимальное число соседей и априорные вероятности классов оцениваются на основании обучающей выборки.

Основная суть метода заключается в том, что для формализации понятия сходства или близости вводится метрика $\rho(x, x')$ в пространстве объектов X, которая, по сути, является функцией расстояния между прецедентами. В качестве такой функции расстояний, в частности, может выступать евклидова метрика.

Для произвольного объекта u из X элементы обучающей выборки $X_\ell = \{x_1, \ldots, x_\ell\}$ располагаются в порядке возрастания расстояний до u:

$$\rho(u, x_{1,u}) < \rho(u, x_{2,u}) < \ldots < \rho(u, x_{\ell,u}),$$

где $x_{i,u}$ — i-й сосед объекта u. Аналогичное обозначение вводится и для ответа на i-м соседе: $y_{i,u} = y(x_{i,u})$. Таким образом, каждый объект u из X порождает свою перенумерацию выборки $X_\ell = \{x_{1,u}, \ldots, x_{\ell,u}\}$.

Простейшим случаем данного метода является, т. н. алгоритм ближайшего соседа (nearest neighbor, NN), обозначим алгоритм через a. Он относит классифицируемый объект u к тому классу, которому принадлежит ближайший объект из обучающей выборки:

$$a(u; X_\ell) = y_{1,u}.$$

Таким образом, распознавание сводится к ранжированию объектов обучающей выборки по степени близости к распознаваемому объекту в соответствии с метрикой ρ. Качество классификации, соответственно, определяется тем, насколько удачно выбрана эта метрика.

Алгоритм k ближайших соседей рассматривает некоторую ближайшую окрестность V_k в признаковом пространстве, содержащую k соседних прецедентов.

Каждый из соседей $x_{i,u}$, $i = 1, \ldots, k$ голосует за отнесение объекта u к классу $y_{i,u}$. В результате объект u относится к тому классу, которому принадлежит большинство из k ближайших к нему объектов обучающей выборки:

$$a(u; X_\ell, k) = \arg\max_{y \in Y} \sum_{i=1}^{k} [y_{i,u} = y].$$

Алгоритм имеет параметр k, который при решении задач конструирования неорганических соединений подбирается по критерию скользящего контроля, т. е. выбирается то значение k, при котором число ошибок классификации минимально:

$$Q(k; X_\ell) = \sum_{i=1}^{k} [a(x_i; X_\ell \setminus \{x_i\}, k) \neq y_i] \to \min_k.$$

В случае разной представительности классов могут применяться модификации метода, использующие различные весовые коэффициенты для разных классов. Достоинствами методов типа kNN являются простота реализации и возможность введения различных модификаций; возможность интерпретации классификации неизвестных объектов путем предъявления ближайшего прецедента или нескольких ближайших прецедентов.

К основным недостаткам метода стоит отнести снижение его эффективности при малых объемах обучающей выборки и высокой размерности признакового пространства. С точки зрения создания классифицирующих

правил метод так же является слабым, так как он вообще не создает каких-либо моделей или правил, а при выборе решения основывается на всем массиве доступных данных обучающей выборки. В связи с этим возникает проблема выбора прецедентов, из которых необходимо формировать обучающую выборку для «хорошей» классификации или прогноза.

Алгоритмы распознавания, основанные на принципе частичной прецедентности

Принцип классификации, применяемый в этих алгоритмах, основан на отнесении распознаваемого объекта *u* к тому классу, в котором имеется большее число «информативных» фрагментов эталонных объектов («частичных прецедентов»), приблизительно равных соответствующим фрагментам объекта *u*. Вычисляются близости — «голоса» (равные 1 или 0) распознаваемого объекта к эталонам некоторого класса по различным информативным фрагментам объектов класса. Данные близости («голоса») суммируются и нормируются на число эталонов класса. В результате вычисляется нормированное число голосов, или оценка объекта *u* за класс $K_j = \Gamma_j(u)$ — эвристическая степень близости объекта *u* к классу K_j.

После вычисления оценок объекта за каждый из классов, осуществляется отнесение объекта к одному из классов с помощью некоторого порогового решающего правила [192]. Решающее правило — это правило (алгоритм, оператор), относящее распознаваемых объект по вектору оценок $(\Gamma_1(u), \Gamma_2(u), ..., \Gamma_k(u))$ в один из классов K_j, или вырабатывающее для объекта «отказ от распознавания». Отказ является более предпочтительным вариантом решения в случаях, когда оценки объекта малы за все классы (объект является принципиально новым, аналоги которого отсутствуют в обучающей выборке), или он имеет две или более близкие максимальные оценки за различные классы (объект лежит на границе классов).

Алгоритмы вычисления оценок (АВО)

Идеи распознавания по частичным прецедентам обобщены в моделях распознавания, основанных на вычислении оценок. Для этих алгоритмов объекты существуют одновременно в самых разных подпространствах пространства признаков. Поскольку не всегда известно, какие сочетания признаков наиболее информативны, то в АВО степень сходства объектов вычисляется при сопоставлении всех возможных или определенных сочетаний признаков, входящих в описания объектов [194].

Используемые сочетания признаков называются опорными множествами или множествами частичных описаний объектов. Вводится понятие обобщенной близости между распознаваемым объектом и объектами обучающей выборки, называемыми эталонными объектами. Эта близость представляется комбинацией близостей распознаваемого объекта с эталон-

ными объектами, вычисленных на множествах частичных описаний. Таким образом, можно сказать, что АВО является расширением метода k ближайших соседей, в котором близость объектов рассматривается только в одном заданном пространстве признаков. В АВО задача определения сходства и различия объектов формулируется как параметрическая и выделен этап настройки АВО по обучающей выборке, на котором подбираются оптимальные значения параметров. Критерием качества служит минимизация ошибок распознавания.

Теоретические возможности АВО, по крайней мере, не ниже возможностей любого другого алгоритма распознавания образов, так как с помощью АВО могут быть реализованы многие операции с исследуемыми объектами. Но расширение возможностей влечет за собой большие трудности для их практического воплощения, особенно на этапе настройки алгоритмов данного типа. Отметим, что трудности, отмеченные при обсуждении метода *k* ближайших соседей, в случае с АВО многократно возрастают.

1.2.2.3. Методы обнаружения логических закономерностей в данных

Для данных методов интерес представляет информация, заключенная не только в отдельных признаках, но и в сочетаниях значений признаков. Они вычисляют частоты комбинаций простых логических событий в подгруппах данных. На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления различных ассоциаций в данных для классификации и прогнозирования. Результат работы данных методов оформляется в виде так называемого дерева решений или правил типа «ЕСЛИ … ТО …».

В целом популярность логических методов обнаружения закономерностей определяется наглядностью результатов их работы. Проблемами являются сложность перебора вариантов за приемлемое время и поиск оптимальной композиции выявленных правил.

Алгоритмы голосования по логическим закономерностям

Математической основой алгоритмов голосования по логическим закономерностям является комбинация логических, оптимизационных и статистических методов [198, 199].

Логический анализ состоит в поиске сходства специальных фрагментов объектов обучения, описанных в терминах признаков. Такое соседство фрагментов считается «типичным» для некоторых классов и «нетипичным» для других.

Подход оптимизации состоит во введении различных числовых критериев для оценки соседства фрагментов и решения математических и программных

проблем для поиска оптимального соседства признаков. Оптимальные решения интерпретируются как логическая закономерность.

Статистические идеи используются введением различных критериев оптимизации (функционалов) и созданием решающих правил в алгоритмах распознавания образов.

Предикат $L_j(x)$ называется логической закономерностью класса K_j при выполнении следующих условий:

$L_j(x_i) = 1$ хотя бы для одного x_i из класса K_j (1);

$L_j(x_i) = 0$ для всех объектов обучающей выборки, не принадлежащих классу K_j, т. е. для $x_i \notin K_j$ (2);

$f(L_j) = max$, где f — некоторый оптимизационный критерий (3).

Критерием качества является функционал:

$f(L_j) = <$*количество объектов обучающей выборки* x_i *из* K_j *:* $L_j(x_i) = 1> / | K_j |$

Предикат $L_j(x)$ называется частичной логической закономерностью класса K_j, если выполнены условия 1 и 3, а условие 2 заменено на более слабое: $(\{x_i \notin K_j | L(x_i) = 1\}) / |\{L(x_i) = 1\}| < \delta$.

Геометрической интерпретацией (рис. 1.2.1) логической закономерности $L_j(x_i)$ является гиперпараллелепипед, содержащий максимальное число объектов обучения из класса K_j, но не содержащий объектов других классов.

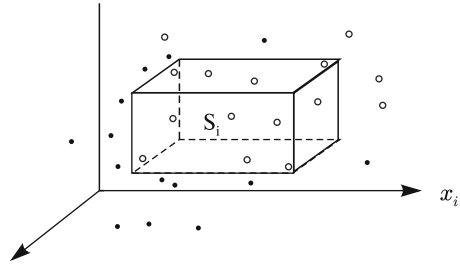


Рис. 1.2.1. Геометрическая интерпретация логической закономерности

Отметим, что АВО и алгоритм голосования по логическим закономерностям можно также выделить в одну группу моделей голосования.

1.2.2.4. Методы, основанные на принципе разделения

Решающее правило в этих алгоритмах основывается на построении поверхности в n-мерном пространстве признаков (гиперповерхности или

набора гиперповерхностей), которая в некотором смысле наилучшим образом будет разделять наборы классов в этом признаковом пространстве. Особо стоит отметить, что такие методы не всегда находят применение в задачах распознавания образов, поскольку построение разделяющих поверхностей с увеличением количества признаков становится в вычислительном плане громоздким.

Линейный дискриминант Фишера

Линейный дискриминант Фишера (Fisher's linear discriminant (FLD)) был первоначально разработан [33] для решения проблемы классификации для случая двух классов. Основная идея метода заключается в проекции векторов признаков на некоторую прямую, что эквивалентно вычислению линейной комбинации их компонент. Сама прямая (коэффициенты линейной комбинации) выбирается таким образом, чтобы отношение расстояния между проекциями средних векторов различаемых классов к сумме разброса проекций векторов внутри каждого класса было максимально. Таким образом, данный метод переводит многомерное пространство признаков в одномерное. Известны способы распространения метода FLD на число классов более двух.

Линейная машина

Этот метод принадлежит к классу методов построения линейных разделяющих гиперплоскостей. Задача при построении такой поверхности состоит в вычислении некоторой линейной относительно признаков функции $f(x) = a_1x_1 + a_2x_2 + \ldots + a_nx_n + a_{n+1}$.

Рассмотрим случай с двумя классами. При классификации используется следующее решающее правило: если $f(u) > 0$, то объект u относится к первому классу, если $f(u) < 0$, то ко второму, а если $f(u) = 0$, то отказ от классификации объекта.

Основной задачей является поиск такой функции $f(x)$, для которой число невыполненных неравенств в системе:

$$Af(u_i) > 0,\ i = 1, \ldots, m_1$$
$$Af(u_i) < 0,\ i = m_1, \ldots, m$$

является минимальным (m — количество объектов). Если система совместна, то достаточно найти любое ее решение $a_1, a_2, \ldots, a_n, a_{n+1}$, если же она несовместна, то находится некоторое «обобщенное» решение, т. е. решение некоторой ее максимальной совместной подсистемы. В результате находится специальная кусочно-линейная поверхность, правильно разделяющая максимальное число элементов обучающей выборки. Другим развитием метода построения линейной поверхности является построение кусочно-линейных поверхностей с помощью метода комитетов [192].

Метод опорных векторов

Метод опорных векторов (support vector machine, SVN) позволяет строить оптимальные линейные или нелинейные разделяющие поверхности [28, 34]. В качестве примера рассмотрим задачу дихотомии, т. е. разбиения объектов обучающей выборки на два непересекающихся класса. Объекты описываются n-мерными векторами

$$x_i = (x^1, \ldots, x^n),\ x_i \in R^n,\ Y = \{-1, +1\}.$$

Строится линейный пороговый классификатор:

$$a(x) = \operatorname{sign}(\sum_{j=1}^{n} w_j x^j - w_0) = \operatorname{sign}(\langle w, x\rangle - w_0),$$

где $x_i = (x^1, \ldots, x^n)$ — признаковое описание объекта x; вектор $w = (w_1, \ldots, w_n) \in R^n$ и скалярный порог $w_0 \in R$ являются параметрами алгоритма. Таким образом, уравнение $\langle w, x\rangle = w_0$ описывает гиперплоскость, разделяющую классы в пространстве R^n.

Предполагается, что выборка линейно разделима, то есть существуют такие значения параметров w, w_0, при которых функционал числа ошибок

$$Q(w; w_0) = \sum_{i=1}^{l} [y_i(\langle w, x_i\rangle - w_0) < 0]$$

принимает нулевое значение. Но тогда существуют и другие положения разделяющей гиперплоскости, реализующие разбиение выборки, т. е. разделяющая гиперплоскость не единственна (рис. 1.2.2). Идея метода заключается в том, чтобы правильным образом распорядиться этой свободой выбора. Необходимо, чтобы разделяющая гиперплоскость максимально далеко отстояла от ближайших к ней точек обоих классов. Первоначально данный принцип классификации возник из эвристических соображений: вполне очевидно, что увеличение зазора (margin) между классами должно способствовать более точной классификации.

Таким образом, с учетом нормировки сформулировано следующее условие для каждого объекта обучающей выборки x_i, $i = 1, \ldots, l$:

$$\langle w, x_i\rangle - w_0 = \begin{cases} \le -1, & \text{если } y_i = -1, \\ \ge 1, & \text{если } y_i = +1. \end{cases}$$

Условие $-1 < \langle w, x\rangle - w_0 < 1$, задает полосу, разделяющую классы. Ни один из объектов обучающей выборки не может лежать внутри этой полосы, границами которой являются две параллельные гиперплоскости с направляющим вектором w. Объекты, ближайшие к разделяющей гиперплоскости, лежат в точности на границах полосы. При этом сама разделяющая гиперплоскость проходит ровно посередине полосы (рис. 1.2.2). Построение оптимальной разделяющей гиперплоскости сводится к решению задачи квадратичного программирования.

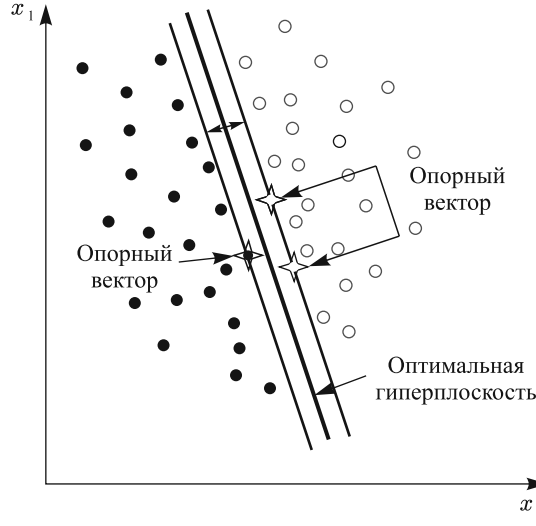


Рис. 1.2.2. Полоса, разделяющая классы

При обобщении метода на случай линейной неразделимости в качестве замены скалярному произведению векторов w и x вводятся «ядровые» функции. В качестве ядровых функций обычно применяются:

- полином скалярного произведения $K(u, v) = (\langle u, v\rangle + 1)^d$;
- гауссиана $K(u, v) = \exp(-\beta||u - v||^2)$;
- гиперболический тангенс $K(u, v) = \text{th}(k_0 + k_1\langle u, v\rangle)$.

Основными достоинствами метода являются широкие возможности для его настройки на конкретную прикладную задачу и сходимость к глобальному максимуму функционала качества за конечное число шагов. Метод опорных векторов оказывается особенно полезным при решении реальных задач на выборках среднего объема с плохо отделимыми в исходном признаковом пространстве классами. Принцип оптимальной разделяющей гиперплоскости приводит к максимизации ширины разделяющей полосы между классами, следовательно, к более уверенной классификации.

Недостатками являются неустойчивость по отношению к шуму в исходных данных. Если обучающая выборка содержит выбросы, они будут учтены при построении разделяющей гиперплоскости. Недостатком этого метода является также сложность выбора параметров, и то, что для классификации используется не все множество образцов, а лишь их небольшая часть, которая находится на границах областей. Помимо этого очень часто

при практическом применении метода опорных векторов наблюдается эффект «переобучения», что естественно снижает перспективность его использования.

1.2.2.5. Нейросетевые алгоритмы

Искусственные нейронные сети (ИНС) базируются на той или иной упрощенной математической модели биологических нейронных систем [27]. Нейронная сеть организует свою работу путем распределения процесса обработки информации между нейроэлементами, связанными между собой посредством синаптических связей. Выявление взаимосвязей в данных осуществляется путем обучения ИНС, в процессе которого осуществляется корректировка весов нейронов.

Нейронные сети являются универсальным аппаратом для задания алгоритмов, т. к. можно использовать нейроны с различными функциями состояния и активации, двоичными, целочисленными и другими значениями весов и входов. В ходе прогнозирования ИНС относит к той или иной области каждый новый объект, поданный на вход сети в виде совокупности значений признаков.

Одним из недостатков использования ИНС в распознавании образов является большое время обучения сети, а также сложность подбора оптимальной архитектуры нейросети для решения конкретных задач. В связи с этим в последние годы применяют модели коллективов нейросетей, в рамках которых несколько нейронных сетей объединяются и используются совместно для решения задачи [35].

Другим недостатком нейросетей является необходимость иметь очень большой объем обучающей выборки. При этом даже обученная нейронная сеть представляет собой «черный ящик». Закономерности, зафиксированные как веса нескольких сотен межнейронных связей, не поддаются анализу и интерпретации человеком.

Нейросети сравнительно легко позволяют найти классифицирующие закономерности в больших объемах данных и получить хорошие результаты классификации в случае больших обучающих выборок. Нейронные сети не столь чувствительны к нарушению гипотезы компактности, т. к. запоминают не границы классов, а области классов. Процесс прогнозирования с использованием обученной нейросети происходит достаточно быстро.

1.2.2.6. Растущие пирамидальные сети

Растущие пирамидальные сети были разработаны в Институте кибернетики им. В. М. Глушкова АН УССР под руководством профессора В. П. Гладуна и более тридцати лет успешно применяются химиками для решения задач прогнозирования [1].

Пирамидальной сетью называется ациклический ориентированный граф, в котором нет вершин, имеющих одну заходящую дугу. Как и в случае нейронных сетей, растущие пирамидальные сети описывают не границы классов, а области объектов, принадлежащие к определенным классам — *объемы понятий*. Понятие в этих методах искусственного интеллекта рассматривается с философской точки зрения — как некоторое обобщение класса объектов в терминах их существенных признаков. *Формирование понятий* может интерпретироваться как процесс поиска закономерностей, свойственных группам объектов.

Первый этап процесса формирования понятий завершается построением пирамидальной сети, представляющей описания объектов обучающей выборки. Сочетания признаков, выделенные на первом этапе, представляют собой «заготовки», из которых формируется логическая структура понятия на втором этапе. Доказано, что алгоритм является сходящимся для понятий любой сложности [186, 187]. Реализация процесса формирования понятий в пирамидальной сети позволяет избежать больших переборов информации, в результате чего появляется принципиальная возможность проводить анализ больших объемов данных.

В результате работы алгоритма в признаковом пространстве строится область для каждого из формируемых понятий, содержащая все точки, представляющие те объекты обучающей выборки, которые входят в объем понятия, и не содержащая ни одной из точек, представляющих другие объекты обучающей выборки.

Существует аналогия между нейронными сетями и растущими пирамидальными сетями. Очень важно, что структура пирамидальной сети формируется автоматически в зависимости от входных данных, а не задается исследователем, как в случае нейросетей. Таким образом, пирамидальная сеть по сути является сетевой памятью, автоматически настраиваемой на структуру входных данных. В результате достигается оптимизация представления информации за счет адаптации структуры сети к структурным особенностям входных данных. Причем, в отличие от нейросетей, эффект адаптации достигается без введения априорной избыточности сети. Возможность интерпретации взаимосвязей в растущих пирамидальных сетях позволяет отнести их к классу семантических сетей, т. е. к структурам данных, состоящим из узлов, соответствующих понятиям, и связей, указывающих на взаимосвязи между узлами [188].

1.2.3. Способы повышения достоверности прогнозов

Как очевидно из краткого обзора методов распознавания образов, в настоящее время не существует универсального подхода к распознаванию,

дающего всегда лучшие результаты. Каждый из методов использует некоторую часть из множества общеизвестных метрик, функций близости, критериев оптимальности, методов оптимизации, способов выбора начальных приближений, способов работы с разнотипными признаками и т. д. и т. п. Основные проблемы практического применения методов распознавания связаны с трудоемкостью и многоэкстремальностью возникающих оптимизационных задач, сложностью сравнения и интерпретации решений, полученных различными методами прогнозирования.

В ситуациях, когда при решении одной и той же задачи распознавания различными алгоритмами находится множество существенно отличающихся решений, перспективным направлением исследований является разработка методов синтеза коллективных решений. Коллективные подходы для решения задач распознавания позволяют некоторым образом объединить разнотипные алгоритмы распознавания и находить оптимальные коллективные решения, в которых компенсируются неточности каждого из используемых базовых методов [36].

Решение задач распознавания коллективами методов

В восьмидесятых годах прошлого века были известны только наиболее простые методы комбинирования алгоритмов, такие как взвешенное голосование и комитетные системы [196]. Практически, эти работы оставались на уровне эвристических приемов и не носили характера научной теории. Один из примеров — голосование результатов распознавания различных алгоритмов, которое относило распознаваемый объект к тому классу, за который проголосовало большинство алгоритмов. В настоящее время известно несколько способов конструирования коллективных решений, наиболее общая теория алгоритмических композиций разработана в алгебраическом подходе к построению корректных алгоритмов, предложенном академиком Ю. И. Журавлёвым и активно развиваемом его учениками [192, 195].

Одной из наиболее простых и естественных идей построения коллективного решения является «объединение» результатов распознавания несколькими алгоритмами в «комитетных» конструкциях. Дело в том, что большинство комитетных методов использует оценки апостериорных вероятностей принадлежности объекта к классу, полученные с помощью исходных алгоритмов распознавания. Часто используется уже упомянутое голосование по большинству, из других методов наиболее употребительными являются методы усреднения, определения минимума, максимума или произведения оценок апостериорных вероятностей.

В методе, предлагаемом Л. А. Растригиным, пространство объектов разбивается на **области компетентности**, и для каждой области строится свой алгоритм [197].

Метод **бустинга** (boosting), предложенный Й. Фройндом и Р. Шапире [200, 201], является алгоритмом усиления классификаторов, путем объединения их в комитет. По сути, алгоритм является разновидностью взвешенного голосования, при этом базовые распознающие алгоритмы строятся последовательно, а процесс их построения управляется следующим образом. Для каждого исходного распознающего алгоритма, начиная со второго, веса обучающих объектов пересчитываются так, чтобы он точнее настраивался на тех объектах, на которых ошибались предыдущие базовые распознающие алгоритмы. Веса алгоритмов также вычисляются, исходя из числа допущенных ими ошибок. Обобщающая способность бустинга исследована, пожалуй, наиболее хорошо. Во многих случаях экспериментально наблюдается почти неограниченное улучшение качества обучения при наращивании числа алгоритмов в полученной композиции [202].

В методе **баггинга** (bagging), предложенном Л. Брейманом [203, 204], производится взвешенное голосование базовых алгоритмов, обученных на различных подвыборках объектов, либо на различных частях признакового описания объектов. При этом выделение подмножеств объектов и/или признаков производится, как правило, случайным образом.

Эмпирические исследования по сравнительному анализу обобщающей способности бустинга и баггинга на четырех реальных задачах показывают, что бустинг работает лучше на больших обучающих выборках, баггинг — на малых [205]. При увеличении длины выборки бустинг повышает разнообразие классификаторов лучше, чем баггинг. Бустинг лучше воспроизводит границы классов сложной формы.

Рассмотрим подробнее алгебраический подход к построению корректных алгоритмов, предложенный академиком Журавлёвым. Данный подход базируется на понятии алгоритмической композиции. Наряду с пространством объектов обучающей выборки X и множеством классов Y, вводится вспомогательное множество R, называемое пространством оценок. Рассматриваются алгоритмы, имеющие вид суперпозиции

$$a(x) = C(b(x)),$$

где функция $b : X \to R$ называется алгоритмическим оператором, функция $C : R \to Y$ — решающим правилом (рис. 1.2.3). Многие алгоритмы классификации имеют именно такую двухстадийную структуру: сначала вычисляются оценки принадлежности объекта к классам, затем решающее правило переводит эти оценки в наименование класса. Значение оценки, как правило, характеризует степень уверенности классификации. В одних алгоритмах это вероятность принадлежности объекта заданному классу, в других — расстояние от объекта до разделяющей поверхности. Возможны и другие интерпретации оценок.

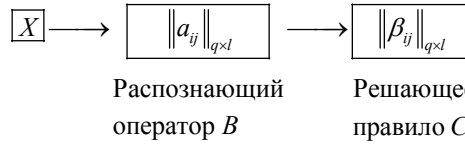


Рис. 1.2.3. Каждый алгоритм распознавания представим в виде произведения распознающего оператора и решающего правила

Алгоритмической композицией, составленной из операторов b_t: $X \to R$, $t = 1, \ldots, T$, корректирующей операции $F : R^T \to R$ и решающего правила $C : R \to Y$ называется алгоритм a: $X \to Y$ вида $a(x) = C(F(b_1(x), \ldots, b_T(x)))$, $x \in X$. Функции $a_t(x) = C(b_t(x))$ называются **базовыми алгоритмами**, $t = 1, \ldots, T$.

Суперпозиции вида $F(b_1, \ldots, b_T)$ являются отображениями из X в R, то есть алгоритмическими операторами.

Вообще говоря, коллективный метод распознавания образов всегда рассматривается, как новый алгоритм распознавания, являющийся некоторой суперпозицией имеющихся алгоритмов. Для получения коллективного решения достаточно задать функцию вычисления оценок апостериорных вероятностей принадлежности распознаваемых объектов к классам.

При этом обычно вводят ограничение, называемое условием согласованности, суть его сводится к тому, что коллективный алгоритм не должен относить распознаваемый объект к классу, к которому его не отнес ни один из исходных алгоритмов. Заметим, что условие автоматически выполняется для задачи дихотомии. При наличии же более двух классов, строго говоря, возможно получение результатов, не удовлетворяющих условию согласованности.

Простейшим комитетным методом является усреднение оценок за классы:

$$P_A(t \mid x) = \frac{1}{p} \sum_{i=1}^{p} P_{A_i}(t \mid x),$$

где A — полученный алгоритм в виде композиции p алгоритмов $A_1, \ldots, A_p$.

Используются также коммитетный метод взятия максимума оценки принадлежности к данному классу по всем исходным алгоритмам:

$$P_A(t \mid x) \sim \max_{1 \le i \le p} P_{A_i}(t \mid x).$$

Метод взятия минимума оценки:

$$P_A(t \mid x) \sim \min_{1 \le i \le p} P_{A_i}(t \mid x) .$$

Метод произведения оценок принадлежности к классу:

$$P_A(t \mid x) \sim \prod_{i=1}^{p} P_{A_i}(t \mid x) .$$

В трех последних случаях апостериорные вероятности требуют масштабирования, чтобы их сумма по всем классам l давала единицу:

$$\sum_{t=1}^{l} P_A(t \mid x) = 1 .$$

Другая концепция построения комитетных решений, использованная в работе, заключается в использовании решающих правил исходных алгоритмов, вместо оценок принадлежности за классы:

$$P_A(t \mid x) = \frac{1}{p} \sum_{i=1}^{p} I_{A_i}^{t}(x) ,$$

где $I_{A_i}^{t}(x)$ — бинарная величина, индикатор классификации объекта x к классу t алгоритмом A_i.

Хорошо зарекомендовавшим себя на практике методом получения коллективных решений является метод Байеса. В данном случае для построения коллективного решения используются статистические свойства выборки. Предполагается, что отдельные алгоритмы комитета являются попарно-независимыми. Часто это требование не выполняется, поэтому этот алгоритм часто называют методом «наивного Байеса». Итоговая оценка за класс рассчитывается по формуле:

$$P_A(t \mid x) = \prod_{j=1}^{p} P(t \mid x, \arg\max_{1 \le k \le l} P_{A_j}(k \mid x) = s_j) ,$$

где s_j — результат классификации объекта j-м алгоритмом распознавания.

Метод Байеса обладает высокой скоростью работы и, как следствие, может быть использован в случае большого количества алгоритмов, составляющих комитет. Также следует обратить внимание на достаточный объем обучающей выборки для получения адекватных оценок условных вероятностей возникновения классов [192].

Динамический метод Вудса и области компетенции

Основной идеей этой группы методов является нахождение для распознаваемого объекта наилучшего в некотором смысле алгоритма из заданного коллектива. Предполагается, что распознающий алгоритм может работать по-разному в разных точках пространства. В одних областях алгоритм работает «хорошо» и практически не совершает ошибок, в других показывает плохие результаты. Для распознаваемого объекта необходимо определить алгоритм, являющийся наилучшим в окрестности данного объекта, тогда получившийся объединенный алгоритм распознавания будет не хуже наилучшего из исходных классификаторов. Вводятся отображения $D: R^n \to \{1,..., f\}$, ставящее в соответствие каждой точке признакового пространства номер соответствующей подобласти из $\{1, ..., f\}$. Дополнительно вводится отображение $F: \{1,..., f\} \to \{1,..., p\}$, по которому для каждой подобласти осуществляется выбор соответствующего классифицирующего алгоритма. Таким образом, для каждой точки пространства ставится в соответствие конкретный классификатор $E: R^n \to \{1,..., p\}$. В общем виде схема работы полученного алгоритма записывается следующим образом [192]:

$$A(S) = A_{E(S)}(S).$$

Соответствующие области, в которых работает тот или иной алгоритм, называют областями компетенции (т. е. выбранный на данной области алгоритм лучше работает, чем другие). Работа метода существенно зависит от количества заданных пользователем областей компетенции. При слишком большом числе областей компетенции возможны многочисленные неоправданные переключения с метода на метод, приводящие к неустойчивой классификации и деградации коллективного решения. Данный метод отличается высокой скоростью распознавания и относительно небольшим временем обучения [197].

Другим подходом является определение меры компетенции каждого алгоритма в окрестности заданного объекта, например следующим образом:

$$E(S) = \arg\max_{1 \le i \le p} \nu_i(U_\delta(S)),$$

где $U_\delta(S)$ — дельта-окрестность объекта S. Таким образом, учитываются локальные свойства алгоритмов. Одним из вариантов такого подхода является метод Вудса. Мера локальной компетенции алгоритма в точке подсчитывается следующим образом. Для каждого алгоритма определяется класс, к которому он относит рассматриваемый объект. Затем производится подсчет доли правильно распознанных объектов этого класса, ближайших

к данному объекту. Количество ближайших объектов класса, используемых для оценки компетенции, является параметром алгоритма и задается пользователем. Создатели метода рекомендуют использовать для этого порядка десяти объектов [192].

Шаблоны принятия решений

Метод [206] заключается в определении профилей каждого класса (информации о совокупном поведении всех исходных алгоритмов на объектах данного класса) и подсчете расстояния между ними и результатом работы коллектива в пространстве оценок. В данном методе отдельные алгоритмы коллектива рассматриваются не как дополняющие друг друга в различных областях пространства образов, а как конкурирующие между собой. Данный метод считается одним из наилучших методов коллективных решений и обладает устойчивыми характеристиками в большом числе экспериментов.

Выпуклый стабилизатор

Использование нескольких классификаторов для решения одной задачи, вообще говоря, увеличивает надежность результата, делая его менее подверженным переобучению. Поскольку оценка степени перенастройки каждого алгоритма может быть получена только косвенным путем, в качестве таковой используется градиент оценки апостериорной вероятности принадлежности объекта классу. Коллективное решение строится исходя из требования правильной классификации объектов контрольной выборки и требования максимальной устойчивости получившегося классификатора в рассматриваемой точке. Под неустойчивостью алгоритма распознавания A на j-том объекте контрольной выборки называется величина:

$$G_A(y_i \mid \varepsilon) = \sum_{k=1}^{l} \frac{1}{\varepsilon^2} \sum_{i=1}^{d} [P(k \mid y_j + \varepsilon_k e_i) - P(k \mid y_j)]^2 \ ,$$

где e_i — единичный вектор соответствующей координаты, $\varepsilon = \{\varepsilon_1, ..., \varepsilon_l\}$. Как видно, значение неустойчивости алгоритма распознавания определяется величиной сдвига ε. Этот параметр можно интерпретировать как некоторое характерное среднее расстояние, на котором будут находиться объекты k-го класса по отношению к ближайшему объекту контрольной выборки своего класса. Вообще говоря, выбирать значение ε можно как из некоторых априорных предположений, так на основании данных, содержащихся в контрольной выборке. При этом полагают, что

$$\varepsilon_k = \varepsilon < \min_{i \neq j} \rho(y_i, y_j),$$

где $\rho(y_i, y_j)$ — расстояние между соответствующими объектами.

Соответственно, **неустойчивостью алгоритма распознавания** A называется величина, равная сумме неустойчивостей на всех объектах контрольной выборки:

$$G_A(\varepsilon) = \sum_{j=1}^{q} G_A(y_j, \varepsilon).$$

Соответственно, алгоритм распознавания A_1 является более устойчивым, чем алгоритм A_2 при выполнении неравенства $G_{A_1}(\varepsilon) < G_{A_2}(\varepsilon)$. Важно отметить, что нельзя рассматривать устойчивость алгоритмов распознавания в отрыве от их эффективности, т. е. качества работа на контрольной выборке. Действительно, легко построить абсолютно устойчивый алгоритм — достаточно все выходы сделать константами, только при этом алгоритм полностью утратит свою распознающую способность.

Отсюда становится ясным, что неустойчивость в качестве критерия можно использовать, например, при сравнении алгоритмов с равным количеством правильно распознанных объектов обучающей выборки (т. е. алгоритмов дающих, по сути, одинаковый результат). В этом случае, следует выбрать более устойчивый алгоритм.

При построении коллективного алгоритма, называющегося **выпуклым стабилизатором**, итоговое решающее правило получается в виде выпуклой комбинации функций оценок исходных алгоритмов, причем коэффициенты выпуклой комбинации зависят от положения распознаваемого объекта относительно объектов контрольной выборки, локальной эффективности соответствующего исходного алгоритма и его локальной устойчивости. Приведем формальное определение выпуклого стабилизатора. Говорят, что алгоритм распознавания A получен из A_1, ..., A_p путем применения выпуклого стабилизатора, если он представим в виде выпуклой комбинацией распознающих операторов:

$$P_A(t \mid x) = \frac{\sum_{k=1}^{q} v_k(x) P_{AF(k)}(t \mid x)}{\sum_{k=1}^{q} v_k(x)},$$

где F:{1, 2, …, q} → {1, 2, …, p} — некоторая функция, определяющая индекс «наилучшего» алгоритма распознавания для каждого объекта контрольной выборки, а v_k: $R^d \to R$ — весовые функции, обладающие следующими свойствами:

$v_k(x) \geq 0$, для любого $k = 1, 2, \ldots, q$,

$v_k(x) \to 0$, при $\rho(x, y_k) \to \infty$,

$$\frac{v_k(x)}{\sum_{k=1}^{q} v_k(x)} \to 1, \quad \text{при } \rho(x, y_k) \to 0.$$

Доказана теорема, по которой алгоритм распознавания A полученный применением выпуклого стабилизатора к семейству алгоритмов $A_1, \ldots, A_p$ является не менее эффективным самого эффективного алгоритма из этого семейства. Выпуклый стабилизатор эффективно применяется для построения коллективных решений на малых выборках. Требование устойчивости решения позволяет значительно снизить эффект перенастройки на обучающую выборку [207].

Краткие выводы

В главе получены следующие результаты:

- Рассмотрены и проанализированы методы конструирования неорганических соединений.
- Формализована постановка задачи компьютерного конструирования неорганических соединений.
- Выявлены особенности неорганического материаловедения, как предметной области, создающие трудности при использовании математических методов распознавания.
- Рассмотрены основные этапы процесса поиска знаний в базах данных (Knowledge Discovery in Databases).
- Рассмотрены методы распознавания образов как математическая основа для поиска многомерных классифицирующих взаимосвязей в признаковом пространстве свойств компонентов химических соединений.
- Рассмотрены коллективные методы, позволяющие объединить разнотипные алгоритмы распознавания и находить оптимальные коллективные решения, в которых компенсируются неточности каждого из используемых базовых алгоритмов.

Глава 2
Анализ архитектурных особенностей информационных систем по свойствам неорганических веществ

Для того чтобы выработать методику построения интегрированной информационной системы (ИС) по свойствам неорганических веществ и материалов (СНВМ) для электронной промышленности, необходимо рассмотреть текущее состояние и принципы построения ИС в указанной предметной области. Очевидно, что попытка построения интегрированной ИС без учета специфики информационных структур, содержащихся в БД ИС СНВМ, равно как и без их семантического понимания, обречена на провал.

2.1. Обзор ИС СНВМ для электроники

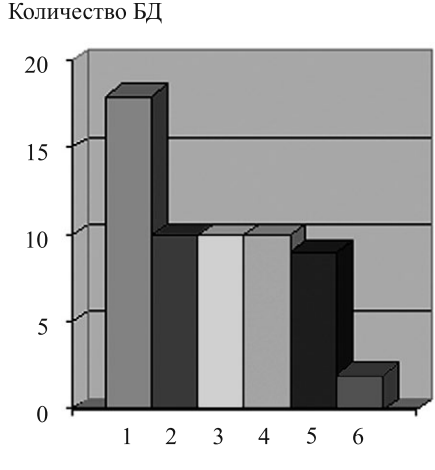
Проблема обеспечения специалистов информацией по свойствам неорганических веществ актуальна для всех промышленно развитых стран. В связи с этим ведется разработка многочисленных информационных систем, основанных на БД по свойствам веществ [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 64–120]. Затраты на создание информационных систем многократно окупаются за счет уменьшения времени на поиск и систематизацию информации и за счет сокращения необоснованного дублирования работ. Рост количества БД, наблюдаемый в последние годы, обусловлен также разработкой мощных и удобных в применении систем управления базами данных (СУБД) и высокопроизводительных компьютеров.

В настоящей работе рассмотрены базы данных по свойствам неорганических веществ, и в первую очередь — БД по свойствам веществ, используемых в электронной промышленности.

Информацию о веществах для электроники можно разбить на три группы: (1) данные о веществах для использования в качестве активных компонентов (полупроводниковых устройств, магнитной памяти, пьезо-

электрических преобразователей, фильтров и гетеродинов, пиро- и сегнетоэлектрических, лазерных, сверхпроводящих, нелинейнооптических, акустооптических, электрооптических устройств и т. д.), (2) информация о веществах для применения в качестве пассивных компонентов (резисторов, трансформаторов, проводников, оптических волокон, печатных плат и т. д.) и (3) данные о вспомогательных веществах (элементоорганических соединениях, кислотах-травителях, пластмассах и т. д.). Помимо этого, большое значение для практических применений имеют сведения о процессах и технологиях получения и обработки веществ и соединений. В табл. 2.1 дан перечень некоторых БД по свойствам неорганических веществ, которые содержат информацию о веществах, используемых в электронике [1].

На рис. 2.1.1 дано распределение наиболее известных баз данных по тематике содержащейся в них информации. Можно заметить, что большинство из них содержат сведения о термохимических, технических и физико-химических свойствах неорганических веществ.



1 Термодинамические или термохимические свойства

2 Технические и технологические свойства

3 Химические и физико-химические свойства

4 Кристаллографические и кристаллохимические свойства

5 Физические (электрические, магнитные, оптические и т. д.) свойства

6 Другие свойства

Рис. 2.1.1. Распределение БД по свойствам неорганических веществ по тематике

Таблица 2.1. Базы данных по свойствам неорганических веществ, используемых в электронике

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«ИВТАНТЕРМ О» — БД термодинамических свойств индивидуальных веществ; «IVTANTHERMO» — DB on thermodynamic properties of individual substances	Объединенный Институт высоких температур РАН (ОИВТ РАН); Joint Institute for High Temperatures of Russian Academy of Sciences (JIHT RAS)	Россия	www.chem.msu.su/rus/handbook/ivtan/a	[83, 220]	Теплофизические и термодинамические свойства (рекомендованные) для неорганических веществ
«ТЕРМАЛЬ» — БД по теплофизическим свойствам чистых веществ; «THERMAL» — DB on thermophysical properties of pure substances	Объединенный Институт высоких температур РАН (ОИВТ РАН); Joint Institute for High Temperatures of Russian Academy of Sciences (JIHT RAS)	Россия	www.thermophysics.ru	[221, 222]	
«ЭПИБИБ» — библиографическая БД по потенциалам взаимодействия и транспортным свойствам разреженных нейтральных газов; «EPIBIB» — documental DB on the intermolecular potentials and transport properties for rarefied neutral gases	Объединенный Институт высоких температур РАН (ОИВТ РАН); Joint Institute for High Temperatures of Russian Academy of Sciences (JIHT RAS)	Россия	www.thermophysics.ru	[223]	БД ориентирована, в первую очередь, на моделирование газотранспортных процессов в микроэлектронике и процессов тепло- и массообмена по газовому тракту энергетических установок

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«NISTTHERMO» — БД по термодинамическим свойствам неорганических и органических веществ; «NISTTHERMO» — DB on thermodynamic properties of inorganic and organic substances	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist103b.cfm; www.nist.gov/srd/nist103a.cfm	[224]	
БД по идеальным газам; NIST/TRC Ideal Gas DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist88.htm	[225]	
«IL Thermo» — БД по ионным жидкостям; «IL Thermo» — NIST Ionic Liquids DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist147.htm	[216]	
«REFPROP» — БД по термодинамическим и транспортным свойствам чистых газов и жидкостей; «REFPROP» — NIST Thermodynamic and Transport Properties of Pure Fluids DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist23.cfm	[226]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
БД по теплофизическим свойствам газов, используемых в полупроводниковой промышленности; DB of the Thermophysical Properties of Gases Used in the Semiconductor Industry	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	http://properties.nist.gov/fluidsci/semiprop/	[216]	
«THERMODATA» — БД по термодинамическим свойствам неорганических веществ; «THERMODATA» — DB on thermodynamic properties of inorganic substances	Ассоциация THERMODATA	Франция	http://thermodata.online.fr/	[85]	Библиография по термодинамическим и теплофизическим свойствам неорганических соединений и сплавов и по фазовым диаграммам
«THERMALLOY» — БД по термодинамическим свойствам неорганических веществ; «THERMALLOY» — DB on thermodynamic properties of inorganic substances	Ассоциация THERMODATA	Франция	http://thermodata.online.fr/theraloy.html	[217]	
«THERMOCOMP» — БД по термодинамическим свойствам неорганических веществ; «THERMOCOMP» — DB on thermodynamic properties of inorganic substances	Ассоциация THERMODATA	Франция	http://thermodata.online.fr/anglais.html	[217]	

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«MTDATA» — БД и пакет программ для расчета фазовых равновесий и термодинамических свойств многокомпонентных систем; «MTDATA» — Software/data package for the calculation of phase equilibria and thermodynamic properties in multicomponent multiphase systems	Национальная физическая лаборатория; National Physical Laboratory (NPL)	Англия	http://www.npl.co.uk/science-technology/advanced-materials/mtdata/	[227, 228]	
DETHERM — БД по теплофизическим свойствам чистых веществ и смесей; DETHERM — DB on Thermophysical Properties of Pure Substances & Mixtures	Общество химической технологии и биотехнологии; Gesellschaft für Chemische Technik und Biotechnologie e.V. (DECHEMA)	Германия	http://i-systems.dechema.de/detherm	[89]	Доступна в сети STN
«TPRC/TPMD» — БД по теплофизическим свойствам; «TPRC/TPMD» — Thermophysical Properties of Matter DB	Информационный центр численного анализа и синтеза Университета Пурдью; Center for Information and Numerical Data Analysis and Synthesis of Purdue University (CINDAS)	США	https://cindasdata.com/	[229]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«SGTE» — БД по термодинамическим свойствам неорганических веществ; «SGTE» — DB on thermodynamic properties of inorganic substances	Европейская научная группа THERMODATA; Scientific Group THERMODATA Europe (SGTE)	ЕС	www.sgte.org/	[97, 98]	
«THERMO-CALC» — Термодинамическая БД и программы для термодинамических расчетов; «THERMO-CALC» — Thermodynamic DB and special software	Корпорация «Thermo-Calc Software»; Thermo-Calc Software Inc.	Швеция	www.thermocalc.com	[99, 230, 231]	
«PPDS» — Термодинамическая БД и программы для термодинамических расчетов; «PPDS» — Software for calculation of physical, thermodynamic and transport properties and, the phase equilibrium of pure components and mixtures and DB on thermophysical properties of fluids (liquids and gases)	Национальная инженерная лаборатория; National Engineering Laboratory (NEL)	Англия	www.ppds.co.uk/	[217]	Программное обеспечение для термодинамических расчетов и БД по теплофизическим свойствам газов и жидкостей

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«MALT2» — БД данных по термодинамическим свойствам индивидуальных веществ с программами расчета равновесного состава; Materials-oriented Little Thermodynamic Database for Personal Computers	Японское общество калориметрии и термического анализа; Japan Society of Calorimetry and Thermal Analysis	Япония	www.kagaku.com/malt	[232, 233]	
«F*A*C*T» — БД по термодинамическим свойствам неорганических веществ и программы для термодинамических расчетов; «F*A*C*T» — DB on thermodynamical properties of inorganic substances (Facility for the Analysis of Chemical Thermodynamics)	Политехническая школа Монреаля; Университет МакЖиль; Центр исследований в вычислительной термохимии; Ecole Polytechnic de Montréal; McGill University; Centre de Recherche en Calcul Thermochimique / Centre for Research in Computational Thermochemistry	Канада	www.crct.polymtl.ca/fact/	[234]	Теплофизические и термодинамические свойства оксидов
«ADAMIS» — БД по свойствам сплавов для микропайки; ADAMIS — Alloy DB for Micro-Solders	Университет Тохоку; Tohoku University	Япония	–	[235, 236, 237]	Термодинамические свойства сплавов для микропайки

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«THERMOSALT» — БД по термодинамическим свойствам смесей расплавленных солей; «THERMOSALT» — DB on thermodynamic properties of molten salts mixtures	Университет Прованса; Universite de Provence	Франция	–	[238]	
«THERSYST» — БД по теплофизическим свойствам сплавов; «THERSYST» — DB on thermophysical properties of solids	Институт ядерной энергии и энергетических систем Университета Штуттгарта Institute for Nuclear Energy and Energy Systems of University of Stuttgart	Германия		[239]	Теплофизические и термодинамические свойства твердых алюминиевых, магниевых и титановых сплавов
БД по термодинамическим свойствам; NIMS Thermodynamic DB	Национальный институт материаловедения; National Institute of Materials Science (NIMS)	Япония	www.nims.go.jp/cmsc/pst/database/periodic.htm	[219]	Термодинамические данные и фазовые диаграммы для двойных систем
«THERM PROP» — БД по термодинамическим свойствам минералов; «THERM PROP» — DB on thermodynamic properties of minerals and related substances	Геологическая инспекция США; Национальный центр термодинамических данных для минералов; U S Geological Survey; National Center for Thermodynamic Data of Minerals	США	–	[217]	

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«NEA-TDB» — БД по теплофизическим свойствам актинидных соединений и других веществ для ядерной энергетики; «NEA-TDB» — Thermochemical DB	Агентство по ядерной энергетике; Nuclear Energy Agency (NEA)	Франция	www.nea.fr/html/dbtdb/	[240]	
«TPDS» — БД по теплофизическим свойствам; Thermophysical Property Data System	Национальный метрологический институт Японии; Национальный институт передовой прикладной науки; Университет Кейо; National Metrology Institute of Japan; National Institute of Advanced Industrial Science and Technology; Keio University	Япония	http://tpds.db.aist.go.jp/index_en.html	[241]	Теплофизические свойства неорганических и органических веществ. Основу БД составляет справочник «Japan Society of Thermophysical Properties; Thermophysical Properties Handbook, 2nd ed.; Yokendo: Tokyo, 2008»
БД по химической кинетике; NIST Chemical Kinetics DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	http://kinetics.nist.gov/kinetics/index.jsp	[216]	Газофазные кинетические данные

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«Диаграмма» — БД по фазовым диаграммам систем с полупроводниковыми фазами; «Diagram» — DB on semiconducting systems phase diagrams	Институт металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН); A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences (IMET RAS)	Россия	http://diag.imet-db.ru/Main.asp	[1, 65, 67, 66, 181, 243]	Фазовые диаграммы систем с полупроводниковыми соединениями, кристаллоструктурные, полупроводниковые и термодинамические свойства фаз
БД по фазовым диаграммам; ACerS-NIST Phase Equilibria Diagrams DB	Национальный институт стандартов и технологий; Американское керамическое общество; National Institute of Standards and Technology (NIST); American Ceramic Society (ACerS)	США	www.nist.gov/srd/nist31.cfm	[216]	Фазовые диаграммы неорганических систем
«Pauling File» — БД по фазовым диаграммам двойных систем и кристаллической структуре двойных соединений; Pauling File on Binary Systems	Японская научно-техническая корпорация; Фирма «Material Phases Data System»; Национальный институт материаловедения; Japan Science and Technology Corporation (JST); Material Phases Data System (MPDS); National Institute for Materials Science (NIMS)	Япония; Швейцария	http://crystdb.nims.go.jp/	[242, 244]	

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«MSIT-PDC» — БД по фазовым диаграммам неорганических систем; MSI-Phase Diagram Centre	Фирма «Materials Science International Services, GmbH» (MSI)	Германия	www.matport.com/phase-diagram-center/buy-online/purchase/selectElements	[245]	
БД по фазовым диаграммам; Alloy Phase Diagram data	Американское общество металлов; ASM (American Society for Metals) International	США	http://www1.asminternational.org/asmenterprise/apd/	[217]	.
БД по фазовым диаграммам; DB on Phase Diagrams	Центр исследований в вычислительной термохимии; Centre de Recherche en Calcul Thermochimique / Centre for Research in Computational Thermochemistry)	Канада	www.crct.polymtl.ca/fact/	[234, 246]	
«TAPP» — БД по фазовым диаграммам и термодинамическим свойствам неорганических веществ; A materials property and phase diagram DB	ESM Software	США	–	[247]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«ICSD» — БД по кристаллической структуре неорганических веществ; «ICSD» — Inorganic Crystal Structure Database	Информационный центр в Карлсруэ; Национальный институт стандартов и технологий; Fachinformationszentrum ***Karlsruhe*** (FIZ Karlsruhe); National Institute of Standards and Technology (NIST)	Германия; США	www.nist.gov/srd/nist84.cfm	[4, 248, 249, 250, 251, 252]	
БД по электронной дифракции; NIST/Sandia/ICDD Electron Diffraction Database	Национальный институт стандартов и технологий; Национальная лаборатория Сандия; Международный центр по дифракционным данным; National Institute of Standards and Technology (NIST); Sandia National Laboratory; International Centre for Diffraction Data (ICDD)	США	www.nist.gov/srd/nist15.htm	[8, 253, 254]	.
«NIST CD» — БД по кристаллическим структурам; «NIST CD» — NIST Crystal Data	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist3.cfm	[255]	
«NSD» — БД по кристаллической структуре металлов, интерметаллидов, сплавов и минералов; «NSD» — NIST Structural DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist83.cfm	[216]	

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«CRYSTMET» — БД по кристаллическим структурам интерметаллических соединений; «CRYSTMET» — Metals and Alloys Crystallographic DB	TOTH — Корпорация информационных систем; TOTH Information Systems, Inc.	Канада	www.tothcanada.com/	[5, 256, 257]	
«Минкрист» — Кристаллографическая и кристаллохимическая БД для минералов и их структурных аналогов; «MINCRYST» — Crystallographic and Crystallochemical DB for Mineral and their Structural Analogues	Институт экспериментальной минералогии РАН (ИЭМ РАН); Institute of Experimental Mineralogy of Russian Academy of Sciences (IEM RAS)	Россия	http://database.iem.ac.ru/mincryst	[258, 259]	
«PDF» — БД по порошковым дифрактограммам неорганических и органических веществ; «PDF» — Powder Diffraction File	Международный центр по дифракционным данным; International Center for Diffraction Data (ICDD)	США	www.icdd.com/	[76, 260]	
«CSD» — БД по кристаллической структуре органических и элементоорганических веществ; «CSD» — Cambridge Structural Database	Кембриджский центр кристаллографических данных; Cambridge Crystallographic Data Centre (CCDC)	Англия	www.ccdc.cam.ac.uk	[261, 262, 263]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«SSD» — БД по структуре поверхности; «SSD» — Surface Structure DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist42.htm	[116]	Данные о структуре поверхности неорганических и органических веществ
«Фазы» — БД по свойствам неорганических соединений «Phases» — DB on inorganic compounds properties	Институт металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН); A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences (IMET RAS)	Россия	http://phases.imet-db.ru/	[1, 118, 181, 243, 267]	
«Кристалл» — БД по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами; «Crystal» — DB on substances with significant acousto-optical, electro-optical and nonlinear-optical properties	Институт металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН); A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences (IMET RAS)	Россия	http://crystal.imet-db.ru/	[1, 65, 72, 181, 243, 269]	

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«Bandgap» — БД по ширине запрещенной зоны неорганических веществ; «Bandgap» — DB on inorganic substances forbidden zone width	Институт металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН); A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences (IMET RAS)	Россия	http://bg.imet-db.ru	[72, 243]	
«Elements» — БД по свойствам химических элементов; «Elements» — DB on chemical elements properties	Институт металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН); A. A. Baikov Institute of Metallurgy and Materials Science of Russian Academy of Sciences (IMET RAS)	Россия	http://phases.imet-db.ru/elements	[243]	
«СМЭТ» — БД по свойствам материалов электронной техники; «SMET» — DB on properties of materials for electronics	Институт неорганической химии им. А. В. Николаева СО РАН (ИНХ СО РАН); A. V. Nikolaev Institute of Inorganic Chemistry of Siberian Branch of Russian Academy of Sciences (NIIC SB RAS)	Россия	–	[65, 273]	Физические, термодинамические, полупроводниковые и кристаллоструктурные свойства полупроводниковых материалов

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
БД по физико-химическим свойствам высокочистых веществ; DB on physical and chemical properties of high-purity substances	Институт химии высокочистых веществ РАН (ИХВВ РАН); Institute of Chemistry of High-Purity Substances of the Russian Academy of Sciences (ICPS RAS)	Россия	–	[65, 84, 274]	Физико-химические свойства высокочистых веществ, определяющие эффективность процессов очистки; информация о коэффициентах распределения примеси при двухфазных равновесиях
«EMIS» — Библиографическая БД полупроводниковым и другим материалам электронной техники; «EMIS» — Electronic Materials Information Service	Институт инженеров-электриков; Institute of Electrical Engineers	Англия	–	[217]	
«MPMD» — БД по свойствам материалов, применяемых в микроэлектронике; «MPMD» — Microelectronics Packaging Materials Database	Информационный центр численного анализа и синтеза Университета Пурдью; Center for Information and Numerical Data Analysis and Synthesis of Purdue University (CINDAS)	США	https://cindasdata.com/	[229]	Данные о физических и физико-химических свойствах материалов, применяемых в микроэлектронике

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«ЭПИДИФ» — БД по коэффициентам взаимной диффузии компонентов газофазной эпитаксии полупроводниковых материалов; «EPIDIF» — DB on Intermolecular Potentials and Diffusion Coefficients for Components of the CVD Processes in Microelectronics	Объединенный институт высоких температур РАН (ОИВТ РАН); Joint Institute for High Temperatures of RAS (JIHT RAS)	Россия	–	[112]	
«WebHTS» — БД по высокотемпературным сверхпроводникам; «WebHTS» — NIST WWW High Temperature Superconductors DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.ceramics.nist.gov/srd/hts/htsquery.htm	[216]	
«SUPERCON» — БД по свойствам сверхпроводников; «SUPERCON» — DB for superconducting materials	Национальный исследовательский институт металлов; National Research Institute for Metals (NRIM) (входит в систему БД National Institute of Materials Science (NIMS))	Япония	http://supercon.nims.go.jp/index_en.html	[275]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
БД по технике высоких магнитных полей и криогенике; High Magnetic Field Engineering and Cryogenics Database	Национальный институт материаловедения; National Institute of Materials Science (NIMS)	Япония	http://yuutsuzu.nims.go.jp/index_eng.html	[219]	Данные о теплопроводности, электропроводности и сверхпроводящих свойствах при низких температурах
«ACDB» — БД по керамикам; «ACDB» — Advanced Ceramics DB	Университет Циньхуа Tsinghua University	Китай	–	[217]	Данные о механических и физических свойствах нитрида кремния и диоксида циркония
«NISTCERAM» — БД по керамикам; «NISTCERAM» — NIST Structural Ceramics DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.ceramics.nist.gov/srd/scd/scdquery.htm	[216]	Данные о теплофизических, механических, коррозионных, свойствах, пористости керамических материалов
БД по функциональным керамическим материалам; Functional Ceramic Materials Database	Лондонский университетский колледж; Лондонский имперал-колледж; University College London, Лондонский университет; Региональная исследовательская лаборатория; Imperial College London, University of London, Regional Research Laboratory	Англия; Индия	http://db.foxd.org	[276]	Данные о диэлектрической проницаемости, ионной диффузии оксидов и других электрокерамических материалов. БД оснащена программами data mining, которые используются для оценки параметров материалов

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«CERAB» — Библиографическая БД по свойствам керамик; «CERAB» — Ceramic Abstracts/World Ceramics Abstracts	Фирма «Cambridge Scientific Abstracts»; Cambridge Scientific Abstracts	США	–	[218]	Доступна в сети STN
БД по инфракрасным спектрам; NIST/EPA Gas-Phase Infrared DB	Национальный институт стандартов и технологий; National Institute of Standards and Technology (NIST)	США	www.nist.gov/srd/nist35.cfm	[216]	
«C and HTC-DATA» — БД по покрытиям и высокотемпературной коррозии; «C and HTC-DATA» — DB on coatings and high temperature corrosion	Университет Прованса; Universite de Provence	Франция	–	[277]	
«АВОГАДРО» — БД для расчетов в области физико-химической газодинамики; «AVOGADRO» — DB for calculations in the field of gaseous dynamics	Институт механики Московского университета; Institute of mechanics of Lomonosov Moscow State University	Россия	–	[278]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«DIPPR» — комплекс БД по свойствам промышленных веществ; «DIPPR» — system of DBs on industrial substance properties	Проектный институт данных о физических свойствах; Design Institute for Physical Property Data (DIPPR)	США	http://dippr.byu.edu/	[90, 91, 264, 265, 266]	Свойства промышленных веществ, зависящие от температуры
«РАДЕН» — БД по радиационным и энергетическим характеристикам двухатомных молекул	Московский государственный университет	Россия	www.elch.chem.msu.ru/cgi-bin/raden/raden.cgi	[279]	
Gmelin-Online Datensystem	Институт неорганической химии Гмелина; Gmelin Institute for Inorganic Chemistry	Германия	–	[280]	Физико-химические свойства неорганических и металлоорганических соединений. Доступна в сети STN
«MOGADOC» — библиографическая БД по свойствам веществ, находящихся в газовой фазе; «MOGADOC» — MOlecular GAsphase DOCumentation DB	Университет Ульма; University of Ulm	Германия	www.uni-ulm.de/strudo/mogadoc/	[101, 281, 282]	Библиография по структурным, электрическим и магнитным свойствам, электронным и микроволновым спектрам молекул неорганических, элементоорганических и органических веществ, находящихся в газовой фазе

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«ASMDATA» — БД ASM по свойствам материалов; «ASMDATA» — ASM Materials Information Online	Американское общество металлов; American Society for Metals International (ASM)	США		[218]	Данные о коррозионных, механических, физических и т. д. свойствах композитов, сталей, металлов, цветных сплавов и пластиков
«COPPERDATA» — библиографическая БД по свойствам меди и ее сплавов; «COPPERDATA» — bibliographic DB of the world's literature on copper, copper alloys and copper technology	Корпорация «Copper Development Association»; Copper Development Association, Inc.	США	www.csa.com/copperdata/	[218]	Данные о механических, электрических, коррозионных и термических свойствах меди и ее сплавов
Информационная система по редкоземельным металлам; Information system of rare earths	Чангчуньский институт прикладной химии Китайской академии; Changchun Institute of Applied Chemistry of Academia Sinica (CIAC)	Китай	–	[108]	Данные о физико-химических свойствах РЗЭ и параметрах процессов экстракции
БД по аморфным материалам; DB on amorphous materials	Университет Тохоку; Tohoku University	Япония	–	[110]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
HYDROGENE DATA — БД по свойствам материалов с водородом; HYDROGENE DATA — DB on properties of materials with hydrogen	Национальный центр научных исследований/Центр изучения химической металлургии; Centre National de la Recherche Scientifique /Centre d'Etudes de Chimie Metallurgique (CNRS/CECM)	Франция	–	[283]	Данные о механических, электрических, электрохимических, магнитных свойствах материалов с водородом, данные по диффузии, растворимости водорода, накопителях водорода, влиянии поверхности, механической и термической обработки, структуре и т. д.
«NRIM CDS» — БД по ползучести; «NRIM CDS» — NRIM Creep Data Sheets	Национальный исследовательский институт металлов; National Research Institute for Metals (NRIM)	Япония	–	[217]	Данные о ползучести, разрыве, прочности сталей и сплавов
«MATADOR» — БД по механическим и физическим свойствам черных и цветных металлов	Объединение моторов и турбин; Motoren und Turbinen Union München GmbH	Германия	–	[217]	

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«DMPME» — БД по механическим свойствам промышленных материалов; «DMPME»–DB of Material Properties for Mechanical Engineering	Шанхайский исследовательский институт материалов; Shanghai Research Institute of Materials	Китай	–	[217]	Механические (включая усталостные), физические, коррозионные и другие свойства черных и цветных металлов, сталей и пластмасс
«AAASD» — БД по свойствам и спецификациям алюминиевых сплавов; «AAASD» — DB on properties and specifications for aluminum alloys and products	Корпорация «Алюминиевая ассоциация»; Aluminum Association, Inc.	США	–	[217]	
«nSOFT FATIMAS MDM» — БД по усталости и трещинообразованию; «nSOFT FATIMAS MDM» — DB on fatigue related data for crack initiation and propagation	Фирма «nCode International»; nCode International Ltd	Англия	www.ncode.com/	[217]	
«CETIM-BDM» — БД по физическим и механическим свойствам промышленных материалов; «CETIM-BDM» — DB on physical and mechanical characteristics of engineering materials	Технический центр промышленности центра информационных технологий; Centre Technique des Industries Mecaniques Centre d'Information Technologique (CETIM)	Франция	–	[217]	Данные о механических и физических свойствах черные и цветных металлов, сталей, композитов

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«SciGlass» — БД по составу и свойствам стекол; «SciGlass» — Glass property DB	Thermex Company; ITC Inc.	Россия	www.sciglass.info	[285, 286, 287, 288]	Данные о механических, оптических, электрических свойствах стекол и фазовых диаграммах
«INTERGLAD» — БД по составу и свойствам стекла; «INTERGLAD» — Glass DB on compositions and properties	New Glass Forum (NGF)	Япония	http://61.194.5.20/interglad6/index.html	[289]	Данные о механических, оптических, электрических свойствах стекол и фазовых диаграммах
БД по прочности материалов; NRIM-JICST Materials Strength Database	Национальный исследовательский институт металлов; Японский информационный центр в области науки и техники National Research Institute for Metals (NRIM); Japan Information Center of Science and Technology (JICST)	Япония	–	[217]	Данные о прочности сталей, сплавов и металлов
«MAT-DB» — БД по механическим и физическим свойствам промышленных сплавов; «MAT-DB» — Materials DB for mechanical and physical properties data of engineering alloys	Объединенный исследовательский центр Европейской комиссии — Институт энергетики; Joint Research Centre of the European Commission — Institute for Energy	Нидерланды	–	[290, 291]	Данные о механических, физических, коррозионных свойствах покрытий, сплавов, керамик

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
БД по свойствам оптических материалов; DB on Optical Materials Properties	Национальные лаборатории Сандия; Sandia National Laboratories	США	–	[217]	
«FASMET» — БД по усталостной прочности металлических материалов; «FASMET» — DB on Fatigue Strength of Metallic Materials	Университет Рицумейкан; Ritsumeikan University	Япония	–	[217]	
«CRAMET» — БД по росту усталостных трещин металлических материалов; «CRAMET» — DB on Fatigue Crack Growth Rates of Metallic Materials	Университет Рицумейкан; Ritsumeikan University	Япония	–	[217]	
«Fatigue DB» — БД по усталости сталей «Fatigue DB» — DB on Fatigue of Steels	Национальный институт материаловедения; National Institute of Materials Science (NIMS)	Япония	https://tsuge.nims.go.jp/top/fatigue.html	[219]	
«MSDRD» — БД по прочности материалов; «MSDRD» — Material Strength DB for Reliability Design of Machines and Structures	Университет Рицумейкан; Ritsumeikan University	Япония	–	[217]	Данные о механических свойствах керамик, композитов, металлов

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«JSSR DB» — БД по пружинам; «JSSR DB» — DB on springs	Национальный исследовательский институт металлов; National Research Institute for Metals (NRIM)	Япония	–	[217]	
«ALFRAC» — БД по механическим свойствам алюминиевых сплавов; «ALFRAC» — Aluminum Fracture Toughness DB	Корпорация «Совет по свойствам материалов»; Корпорация «Алюминиевая ассоциация»; Национальный институт стандартов и технологий; The Materials Properties Council, Inc.; Aluminum Association, Inc.; National Institute of Standards and Technology	США	–	[218]	Данные о прочности на разрыв и излом при плоских деформациях и надрезах 32-х алюминиевых сплавов
«MARTUF» — БД по свойствам сталей, применяемых в судостроении; «MARTUF» — DB on Toughness of Marine Steels	Корпорация «Совет по свойствам материалов» The Materials Properties Council, Inc. (MPC)	США	–	[218]	Данные о механических, коррозионных и прочих свойствах сталей
«MDF» — БД по свойствам сплавов черных и цветных металлов; «MDF» — Metals Data File	Фирма «Cambridge Scientific Abstracts»; Cambridge Scientific Abstracts	США	–	[218]	Доступна в сети STN

Продолжение таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«METALCREEP» — БД по ползучести и разрушающему напряжению алюминиевых сплавов и сталей; «METALCREEP» — DB of creep and rupture stress properties for aluminum alloys and steels	Национальная сеть данных по материалам; National Materials Property Data Network Inc.	США	–	[218]	
«SPTD1» — БД по структурным фазовым переходам; «SPTD1» — DB of Phase Transitions in Crystals with a Single Phase Transition	Институт низкотемпературных и структурных исследований Польской АН; Institute of Low Temperature and Structural Research of PAN	Польша	–	[292]	
«ICSDB» — БД по структуре несоразмерных фаз; «ICSDB» — Incommensurate Structures DB	Университет Па Васко; Universidad del Pais Vasco	Испания	www.cryst.ehu.es/icsdb/about.html		
БД по несоразмерным фазам DB of incommensurate phases	Католический университет Лоувейна; Universite Catholique de Louvain	Бельгия	www.uclouvain.be/en-imcn.html	[293]	

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
«CompoTherm» — БД и система прогноза теплофизических свойств композитов; «CompoTherm» — Thermophysical Property Prediction System for Composite	Национальный институт материаловедения; National Institute of Materials Science (NIMS)	Япония	http://composite.nims.go.jp	[270]	Данные о теплофизических, термодинамических, механических свойствах композитов
Система БД по электронной структуре; Database System for Electronic Structures	Японская научно-техническая корпорация; Национальный институт материаловедения; Japan Science and Technology Corporation (JST); National Institute of Materials Science (NIMS)	Япония	http://caldb.nims.go.jp/	[219]	
«CCT DB» — БД по сварке; «CCT DB» — Welding DB	Японская научно-техническая корпорация; Национальный институт материаловедения; Japan Science and Technology Corporation (JST); National Institute of Materials Science (NIMS)	Япония	http://inaba.nims.go.jp/Weld	[219]	
«Data-Free-Way» — БД по ядерным материалам; «Data-Free-Way» — Nuclear Materials Database	Национальный институт материаловедения; National Institute for Materials Science (NIMS)	Япония	http://dfw.nims.go.jp/	[294, 295, 296]	

Окончание таблицы 2.1

Название БД	Организация	Страна	URL-адрес	Ссылки	Примечания
БД по механическим свойствам; Structural Materials DB	Национальный институт материаловедения; Японский научно-исследовательский институт атомной энергии; Японский институт по переработке ядерного топлива; National Institute of Materials Science (NIMS); Japan Atomic Energy Research Institute (JAERI); Japan Nuclear Cycle Development Institute (JNC)	Япония	http://tsuge.nims.go.jp/	[219]	
Система БД по материалам для сосудов высокого давления; Database System for Pressure Vessel Materials	Национальный институт материаловедения; National Institute of Materials Science (NIMS)	Япония	http://pvmdb.nims.go.jp/index_eng.html	[219]	
БД по диффузии; Diffusion DB	Национальный институт материаловедения; National Institute of Materials Science (NIMS)	Япония	http://diffusion.nims.go.jp/index_eng.html	[219]	
БД по нанокомпозитам; DB on nanocomposites	Российский химико-технологический университет им. Д. И. Менделеева; D. Mendeleev University of Chemical Technology of Russia	Россия		[297]	

В последние годы наблюдается тенденция к кооперации в разработке БД и к интеграции уже созданных БД как на национальном, так и на международном уровнях, в том числе и в рамках CODATA и ЮНЕСКО. Это вызвано стремлением устранить дублирование работ и уменьшить затраты на разработку и поддержание БД. Многие БД доступны в режиме удаленного доступа с использованием телекоммуникационных сетей [2, 3]. Наиболее мощные системы баз данных предлагают NIST и STN [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17].

Следует отметить, что только несколько баз данных, указанных в табл. 1.1, созданы для информационного обеспечения специалистов, разрабатывающих и использующих вещества, используемые в электронной технике: СМЭТ [78, 79, 80, 81], БД по сверхпроводникам, разработанная в Токийском университете [104], SUPCONA [103], NIST/NRIM High Temperature Superconductors Database [6], NISTCERAM [7], ЭПИДИФ [112], БД, разработанная в МИТХТ [79, 113], Microelectronics Packaging Materials Database, предлагаемая информационным центром CINDAS при Purdue University [95].

Базы данных по свойствам веществ для электроники ИМЕТ РАН [66, 70, 72]: БД по фазовым диаграммам систем с полупроводниковыми фазами «Диаграмма» и БД по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами «Кристалл» — проблемно ориентированы на химиков-технологов и специалистов в области электроники. Существенными отличиями разработанных в ИМЕТ РАН информационных систем от созданных ранее являются:

- сбор и экспертная оценка качества данных осуществляются высококвалифицированными специалистами;
- отсутствие аналогов;
- возможность доступа из сети Интернет.

2.2. Создание ИС по информационным ресурсам неорганической химии «IRIC»

В современном мире наблюдается неуклонный рост потоков информации во всех отраслях человеческой деятельности. За последние десятилетия неорганическим материаловедением был накоплен колоссальный массив сведений по широкому спектру свойств современных материалов, а также технологиям их получения. По мере развития научно-технического прогресса наблюдалась естественная эволюция средств доставки информации до потребителей. В современном материаловедении за последние десятилетия был пройден путь от попыток систематизации нако-

пленной информации в справочниках, статьях, монографиях до повсеместного использования специализированных информационных систем, использующих базы данных. На текущий момент именно базы данных в наибольшей степени отвечают потребностям специалистов по неорганическому материаловедению, поскольку обеспечивают быстрый поиск информации, поддерживаемой, в отличие от бумажных носителей, в актуальном состоянии.

В последнее время в мире наблюдается неуклонный рост числа материаловедческих баз данных (БД) и основанных на них информационных систем (ИС). Разработка информационных систем по свойствам неорганических веществ и материалов (ИС СНВМ) на основе БД ведется во всех промышленно развитых странах на многих языках. Среди крупнейших разработчиков ИС СНВМ, как отмечалось ранее, стоит выделить NIST, STN и NIMS.

Несмотря на увеличивающиеся объемы данных, содержащиеся в рамках БД, ни одна из них не содержит полного описания всех свойств веществ. Поэтому всестороннее изучение свойств того или иного материала требует анализа информации из целого ряда информационных систем. Такой анализ является необходимым, поскольку в современных многофункциональных устройствах только исчерпывающая характеристика материала позволяет материаловеду принять решение об его использовании. Таким образом, перед исследователем встает проблема поиска требуемой информации в разрозненных ИС СНВМ, что невозможно без систематизации самих ИС СНВМ. Именно задача систематизации наиболее значимых информационных ресурсов по свойствам неорганических веществ решалась на базе ИМЕТ РАН при разработке ИС «IRIC» по информационным ресурсам в области неорганической химии (IRIC — Information Resources on Inorganic Chemistry) [312].

2.2.1. Схема данных

Как известно, любая ИС состоит наполовину из данных, а наполовину — из программного кода. Схема данных является наиболее критичной частью для реализации любой ИС, поскольку основные функции ИС разрабатываются именно на уровне схемы данных. Таким образом, если схема данных поддерживает некоторую функциональность, то программный код способен реализовать ее. Если нет, то, как бы хороша не была программная реализация, конечная ИС не сможет качественно поддерживать функции, изначально не заложенные в схему БД. Поэтому важно было выделить основные сущности для ИС «IRIC» и отношения между ними, которые позже лягут в основу проектируемой БД [313].

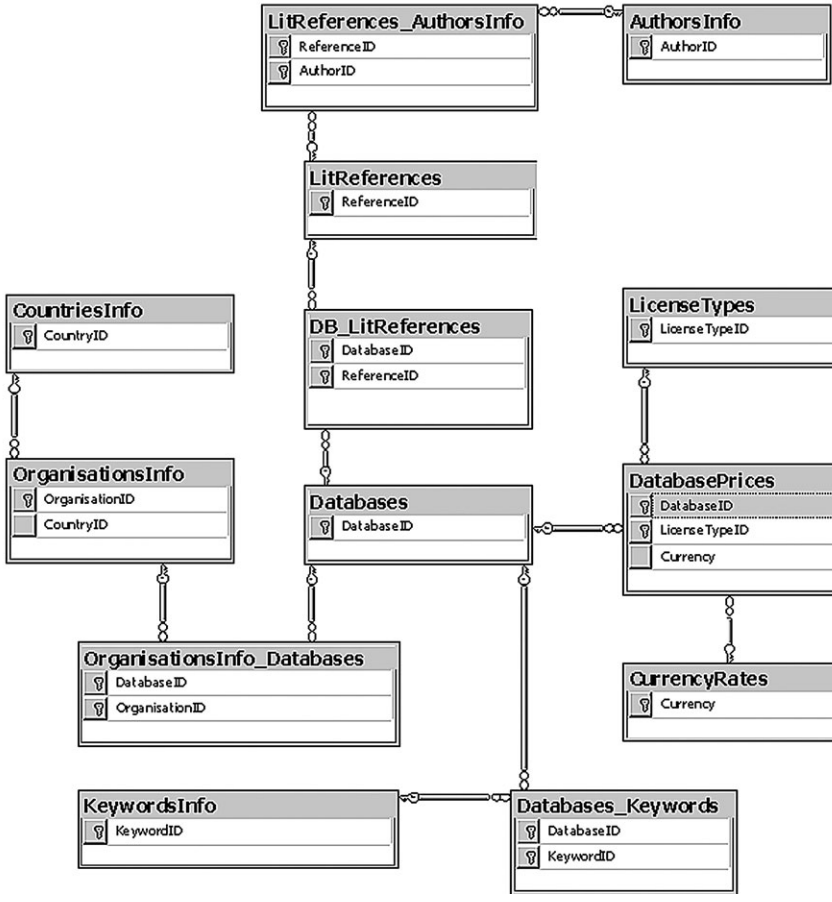


Рис. 2.2.1. Логическая модель ИС «IRIC»

Перечислим основные из сущностей, выделенные при разработке схемы данных: **страны, организации-разработчики, БД, ключевые слова, литературные публикации и их авторы, условия доступа к БД (политики доступа)**.

После составления вербальной модели была составлена ER-диаграмма (диаграмма сущность— связь), которая затем была преобразована в физическую модель данных, представленную в Microsoft SQL Server 2008. Все таблицы создавались с помощью SQL DDL (Data Definition Language) операторов, в которых описывались атрибуты отношений, их типы данных, а так же связи с другими таблицами. Приведем пример для таблицы

(LitReferences), описывающей литературные публикации, из которых стало известно о существовании описываемой той или иной ИС СНВМ:

```
CREATE TABLE [dbo].[LitReferences](
    [ReferenceID] [int] NOT NULL,
    [Article] [varchar](2048) NOT NULL,
    [Source] [varchar](2048) NOT NULL,
    [Year] [int] NOT NULL,
    [Volume] [varchar](32) NOT NULL,
    [Number] [varchar](32) NOT NULL,
    [Pages] [varchar](32) NOT NULL,
CONSTRAINT [PK_References] PRIMARY KEY CLUSTERED
([ReferenceID] ASC) ON [PRIMARY]
) ON [PRIMARY];
```

В результате получили логическую реляционную и физическую модель данных в Microsoft SQL Server 2008 (рис. 2.2.1). Связи на диаграмме связывают одноименные поля соединяемых таблиц, первичные ключи таблиц отмечены знаками «ключ». Следует отметить, что при разработке схемы БД учитывалась возможность представления всей информации на двух языках — русском и английском. Это впоследствии открыло возможность для написания русскоязычного и англоязычного интерфейса к БД «IRIC», что позволило широкому кругу материаловедов не только в нашей стране, но и за рубежом использовать данную ИС.

2.2.2. Web-приложение

Как известно, интернет является средой обеспечивающей быстрый доступ к информации из любой точки мира, поэтому для доступа к информации ИС «IRIC» было разработано Web-приложение, написанное на классическом ASP (Classic ASP) с использованием ActiveX Data Objects (ADO) в качестве интерфейса доступа к разработанной выше БД. Пример скрипта на языке VBScript, открывающий подключение к БД «IRIC», приведен ниже:

```
        <%
        Dim BDN, RSN
        Function Initialize(ConnectionString)
        Set BDN = Server.CreateObject("ADODB.Connection")
        BDN.Open ConnectionString
        Set RSN = Server.CreateObject("ADODB.Recordset")
        RSN.ActiveConnection = BDN
        RSN.CursorLocation = 3
        RSN.CursorType = 0
        RSN.LockType = 1
        End Function
        Call Initialize("Provider = SQLOLEDB;Data
Source = 193.233.10.65;Initial Cata-
log = Iric;UID = xxxxxxxx;PWD = xxxxxxxx; ")
        %>
```

Результатом работы VB-скрипта, является открытое соединение с БД, и объект ADO Recordset, готовый к выполнению поисковых запросов. После завершения работы с соединением, его необходимо закрыть, вызвав метод Close объекта Connection, т. е. в данном случае BDN.Close.

На основании структуры данных в БД был разработан рубрикатор Web-приложения (рис. 2.2.2), впоследствии трансформировавшийся в главное меню и его подразделы (рис. 2.2.3). При создании разделов рубрикатора использовалась созданная с участием автора система SimpleCMS, облегчающая создание навигационных элементов Web-приложений [310].

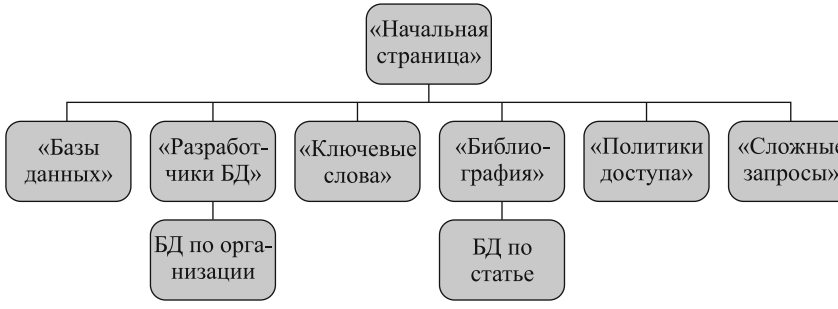


Рис. 2.2.2. Структура Web-приложения «IRIC»

В 2013 году в Web-приложение были встроены поисковые средства Google (Google Search engine), позволяющие осуществлять качественный полнотекстовый поиск по всему содержимому, генерируемому Web-приложением, включая полнотекстовые статьи в PDF-формате [310].

Эти средства наряду с реализованным в «IRIC» поиском ИС СНВМ по ключевым словам дает пользователям дополнительную гибкость при работе с ИС. ИС «IRIC» поддерживает ряд поисковых запросов, которые отличаются количеством и типом критериев поиска. Например, поддерживается поиск по организации-разработчику, литературным публикациям, по заданному набору ключевых слов и др. На рис. 2.2.4. показан фрагмент интерфейса, содержащий поиск по ключевым словам «энтальпия» и «энтропия», как видно из снимка экрана, на сегодняшний день ИС «IRIC» известно три БД, удовлетворяющие критерию поиска [298].

ИС «IRIC» доступна круглосуточно и обеспечивает мгновенный доступ к информации через любую программу-браузер. Русскоязычный интерфейс доступен по адресу http://iric.imet-db.ru, полный англоязычный аналог — по адресу http://en.iric.imet-db.ru. Вся информация в ИС предоставляется в открытом доступе (бесплатно) для всех желающих. В настоящее время «IRIC» содержит сведения о 122 информационных ресурсах, созданных в мире.

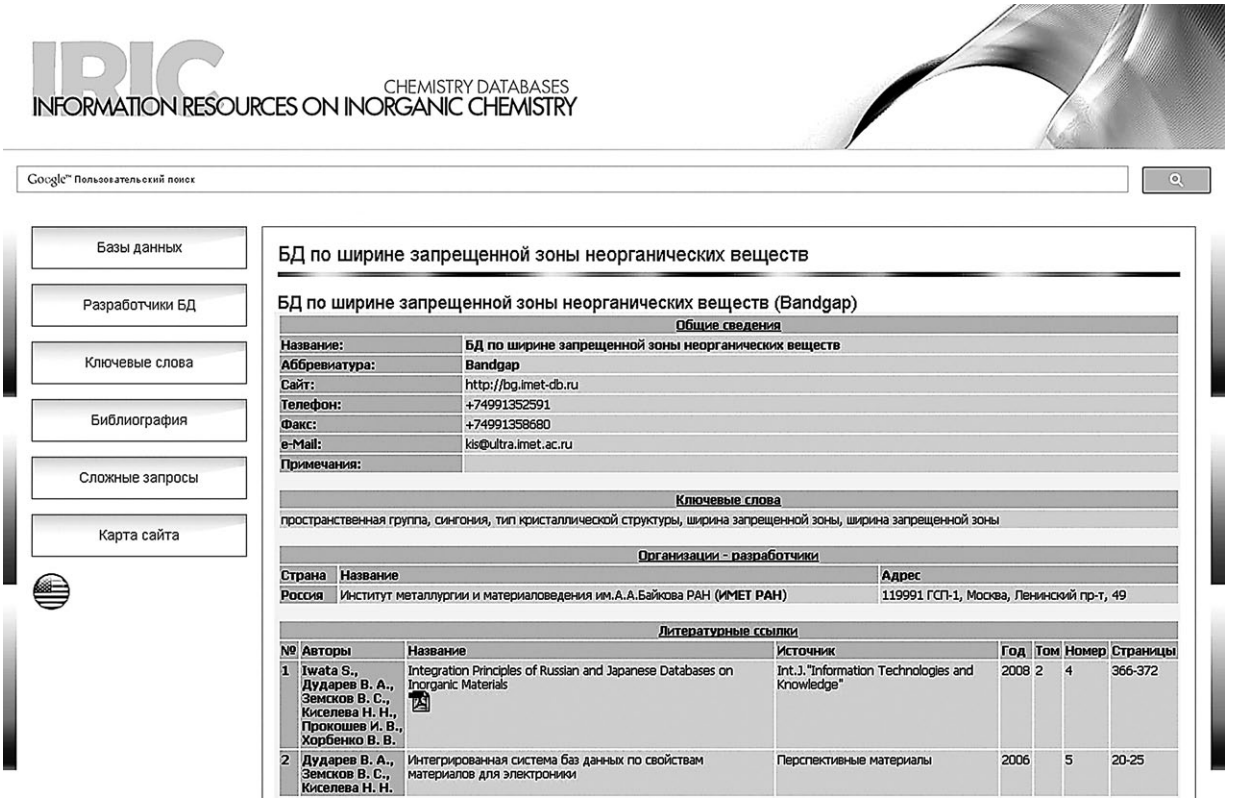


Рис. 2.2.3. Пользовательский интерфейс ИС «IRIC»

Список БД по ключеым словам

Фильтр по ключевым словам: энтальпия, энтропия

№	Название (аббревиатура)	Контакты	Комментарий
1	БД термодинамических свойств индивидуальных веществ **(ИВТАНТЕРМО)** [подробнее...]	**Сайт:** **http://www.chem.msu.su/rus/handbook/ivtan** Тел: +74954851000 Факс: +74954851000 e-Mail: iorish@ihed.ras.ru	Теплофизические (рекомендованные) данные для неорганических веществ.
2	БД по термодинамическим и транспортным свойствам чистых газов и жидкостей **(NISTFLUIDS)** [подробнее...]	**Сайт: http://www.nist.gov/srd/nist23.htm** Тел: +13019752208 Факс: +13019260416 e-Mail: Joan.Sauerwein@nist.gov	Теплофизические и транспортные свойства чистых неорганических и органических жидкостей и газов.
3	БД по термохимическим свойствам органических и неорганических веществ с одним или двумя атомами углерода **(JANAF)** [подробнее...]	**Сайт:** Тел: Факс: e-Mail:	Теплофизические и химические свойства неорганических веществ и органических веществ с одним-двумя атомами углерода.

Всего найдено: 3. 1-3

Рис. 2.2.4. Пример результатов запроса к ИС «IRIC» по ключевым словам

Отличительными особенностями «IRIC» являются:

- первый в мире каталог по информационным ресурсам в области неорганической химии и материаловедения, содержащий библиографические источники;
- интерфейс на русском и английском языках;
- возможность поиска по метаданным и полнотекстового поиска, включая библиографию.

Разработанная ИС «IRIC» позволяет материаловедам не только получать информацию о существующих в мире БД по свойствам неорганических веществ на русском и английском языках, но и проводить поиск таких БД по многокритериальным запросам. Создание ИС «IRIC» позволяет систематизировать имеющеюся информацию в материаловедческих БД на самом верхнем уровне и указать варианты наиболее разумной интеграции созданных ИС с целью минимизации времени затрачиваемого специалистами на поиск требуемой информации. А это, в свою очередь, является важным шагом на пути к созданию единой интегрированной материаловедческой информационной системы следующего поколения.

2.3. Архитектура современных информационных систем по свойствам веществ

2.3.1. Использование трехзвенной архитектуры

При рассмотрении ИС СНВМ, информация по которым доступна в ИС «IRIC», можно констатировать, что большинство их них спроектированы и разработаны с учетом современных требований, предъявляемых к построению ИС. Так, большинство систем располагают Web-интерфейсами и доступны для пользователей через глобальную сеть Интернет. Если рассматривать архитектурно-технологические аспекты построения этих систем, то можно обнаружить много схожих моментов.

При разработке программного обеспечения (ПО) всех интернет-ориентированных ИС использовалась распределенная модель построения приложений на основе парадигмы «клиент-сервер». В соответствии с положениями этой парадигмы реализуется разделение операций по обработке и отображению информации между сервером и персональным компьютером клиента — пользователя информационной системы. При построении ИС использовалась классическая трехзвенная архитектура. Трехзвенная архитектура — вариант архитектуры клиент-сервер, в которой пользовательский

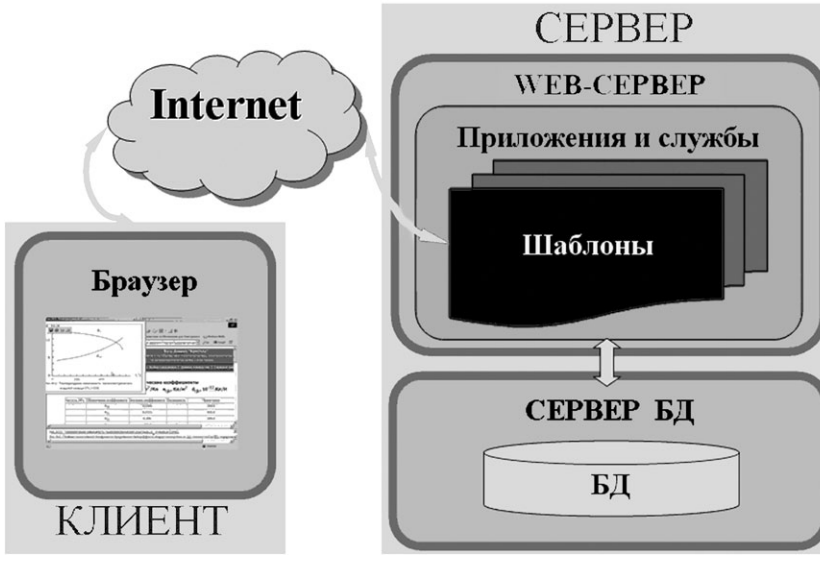


Рис. 2.3.1. Трехзвенная архитектура ИС с доступом пользователей через Интернет

интерфейс, логика работы приложения, доступ к данным и хранение данных разрабатываются и функционируют как независимые модули, зачастую на различных программно-аппаратных платформах (рис. 2.3.1). Стрелками на рисунке отображены потоки информации между звеньями.

Рассмотрим разделение функций между звеньями ИС, построенных по принципам этой архитектуры. Сразу следует отметить, что ИС может быть условно разбита на две составные части: это серверная часть и клиентская часть (на основе парадигмы «клиент—сервер»). При этом взаимодействие частей осуществляется с использованием глобальной сети Интернет на основе стека протоколов TCP/IP [153].

Рассмотрим более подробно структуру клиента. Под клиентом понимается удаленный пользователь информационной системы, взаимодействующий с ней через сеть Интернет. Следует сразу отметить, что поскольку при разработке Интернет-ориентированных ИС ставится задача обеспечения максимально простого и стандартного программного обеспечения на клиентской части, то принимается решение об использовании Интернет-браузера. Интернет-браузер является стандартным компонентом ПО в современных персональных компьютерах и, как правило, входит в состав всех распространенных операционных систем (ОС). Например, во все ОС корпорации Microsoft по умолчанию встраивается браузер Microsoft

Internet Explorer (текущая версия 11). Таким образом, от пользователя ИС не требуется устанавливать дополнительное ПО, обеспечивающее взаимодействие с рассматриваемыми ИС. Строго говоря, пользователю достаточно иметь любой распространенный браузер (Google Chrome, Opera, Mozilla FireFox, Microsoft Internet Explorer, Apple Safari и т. п.) и быть подключенным к сети Интернет. Этот вариант клиентского ПО в литературе также часто называют термином «легкий клиент», подчеркивая тем самым не только простоту клиентского ПО, но и то, что основная масса всех вычислений возлагается на сервер, а клиент лишь получает готовый результат в виде динамически сгенерированных HTML-страниц. В контексте трехзвенной архитектуры программного комплекса клиентское ПО является звеном, ответственным за представление (или визуализацию) обработанной информации для конечного пользователя ИС.

Серверная часть информационных систем ИМЕТ РАН взаимодействует с браузером удаленного клиента через сеть Интернет, используя для этого протокол HTTP [154], являющийся одним из протоколов прикладного уровня стека TCP/IP. Вся основная обработка информации, равно как и ее хранение возлагается на серверную часть ПО (сервер также в таких случаях называют «тяжелым»). Серверная часть может быть в свою очередь условно разбита на два звена — звено обработки информации и звено хранения информации. Следует сразу отметить, что такое деление далеко не всегда бывает условным, т. к. активно используются сложные сценарии развертывания информационных систем. При сложных сценариях развертывания эти два звена выносятся не только логически, но и физически на разные сервера. Такой прием, надо отметить, активно используется в ИМЕТ РАН, для повышения производительности, масштабируемости и устойчивости функционирующих ИС СНВМ.

Рассмотрим кратко звено хранения информации. Когда речь заходит о хранении информации, как правило, под ПО, реализующим эти возможности, понимают сервера баз данных. В настоящее время, наибольшее распространение получили реляционные сервера БД. Зачастую подобные сервера основываются на системах управления базами данных (СУБД), предлагаемых крупнейшими разработчиками ПО. Не исключение и ИМЕТ РАН, где в качестве серверов БД используются СУБД Microsoft SQL Server 2008 x64 (под управлением ОС Windows Server 2008 x64) и Oracle 8*i* (под управлением ОС Sun Solaris 2.5).

В качестве звена обработки информации в современных ИС СНВМ выступает сервер приложений (или Web-сервер, например Microsoft IIS, Apache, nginx). В настоящее время все современные Web-сервера поддерживают средства динамической генерации страниц по запросу пользователя. Это обстоятельство, наряду с богатыми возможностями, предоставляемыми для разработчиков подобным ПО, позволяет рассматривать

современный Web-сервер в качестве сервера приложений. Под этим, прежде всего, подразумевается то, что в современных Web-серверах обеспечивается поддержка одновременного размещения (хостинга) нескольких Web-ресурсов, включающих в себя Web-приложения и Web-сервисы. Притом возможно как размещение Web-приложений в изолированных адресных пространствах в рамках отдельных процессов и даже виртуальных машин, так и обеспечение распределенной работы одного Web-приложения сразу на нескольких Web-серверах (так называемая Web-farm). Это значительно повышает масштабируемость и отказоустойчивость современных Web-приложений. В качестве Web-серверов в ИМЕТ РАН, например, используются сервера на базе технологий Microsoft (это Microsoft Internet Information Services (MS IIS) 7.0). Для интеграции Web-серверов и серверов БД используются широко распространенные технологии клиентского доступа к данным (прежде всего ADO, OLE DB и ODBC). Таким образом, по запросу пользователя ИС Web-сервер осуществляет динамическую генерацию (сборку) HTML-страницы, обращаясь непосредственно к серверу БД за информацией, которую необходимо отобразить.

2.3.2. Недостатки ИС СНВМ

Рассмотренная выше трехзвенная архитектура не описывает полностью работу современных ИС СНВМ. Как отмечалось ранее при разработке ИС IRIC, особую важность имеют информационные структуры, которыми оперирует ИС. В данном исследовании обзор информационных структур проводится через призму возможности использования исследуемых ИС СНВМ для первичной подготовки данных, которые могут быть использованы в комплексах компьютерного конструирования неорганических веществ. Более того, исследование текущего состояния и принципов построения ИС СНВМ с точки зрения информационных структур, содержащихся в БД, и их семантики, является необходимым условием разработки интегрированной ИС. На основе анализа пользовательских интерфейсов крупнейших международных материаловедческих комплексов: AtomWork (бывш. Pauling File, NIMS, Япония), SpringerMaterials (the Landolt-Börnstein Database) были выявлены следующие недостатки в построении ИС, затрудняющие поиск информации и дальнейшее использование ИС в качестве источников данных для программ компьютерного конструирования неорганических соединений:

- **отсутствие функций поиска информации по количественному составу соединения**;
- **отсутствие функций поиска по значениям физико-химических свойств**.

Эти недостатки связаны с тем, что рассмотренные ИС СНВМ по сути являются документальными, с некоторыми возможностями поиска по ключевым словам, а не фактографическими, как требуется в СППР при прогнозировании свойств веществ. Поэтому выявленные недостатки можно квалифицировать как системные проблемы в архитектуре данных ИС СНВМ, препятствующие эффективному поиску материаловедческих данных. Таким образом, использование подобных ИС СНВМ в комплексах компьютерного конструирования неорганических соединений невозможно.

Выявленные архитектурные недостатки были устранены при разработке автором некоторых ИС ИМЕТ РАН, описанных ниже. Отличительной особенностью разработанных ИС СНВМ является возможность поиска данных по сложным запросам: запросы с учетом количественного состава и кристаллических модификаций химических соединений, а так же сложносоставные запросы, связанные с поиском материалов с учетом нескольких значений свойств, например, твердости и температуры плавления. Эти типы запросов, как правило, не реализованы в материаловедческих ИС, разработанных в других организациях, что затрудняет использование подобных ИС при компьютерном конструировании неорганических соединений.

2.3.3. Обобщенная структура данных для ИС СНВМ

При анализе разнородных ИС СНВМ установлено, что описание сущностей и их свойств происходит с разной степенью детализации. Отмечено, что значения свойств, хранимых в разных информационных источниках, определяются, в первую очередь, составом неорганических веществ (набором образующих их атомов и их соотношением). В свою очередь, физические свойства веществ во многом зависят от кристаллической структуры. Анализ информации, содержащейся в ИС СНВМ, присутствующих в каталоге «IRIC», позволил составить иерархию химических понятий (рис. 2.3.2).

Обозначив объекты второго уровня общим термином «вещество» получаем трехуровневую иерархию понятий: система, вещество и кристаллическая модификация. Вся информация о свойствах химических объектов, описываемых в интегрируемых ИС СНВМ, может быть представлена на одном из этих трех уровней. Приведенная иерархия понятий, являясь, по сути, упрощенной онтологией для рассматриваемой в настоящей работе предметной области, имеет особую важность для построения информационных моделей данных как при разработке ИС СНВМ, так и при их интеграции (рис. 2.3.2).

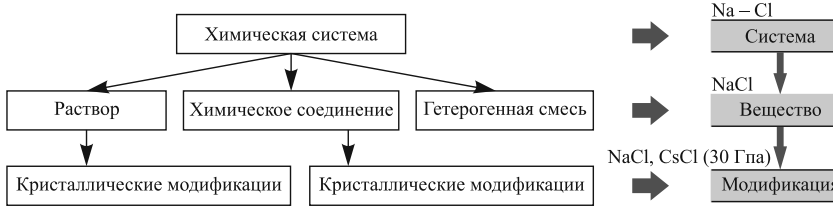


Рис. 2.3.2. Иерархия понятий предметной области

Для строгой формализации предложенной иерархии используется теория множеств. Множество химических систем обозначается S, множество химических веществ — C, а множество кристаллических модификаций — M. Химическая система обозначается s (где $s \in S$), химическое вещество — c (где $c \in C$), а химическая модификация — m (где $m \in M$).

Химическая система s представляется множеством обозначений химических элементов e_i: $s = \{e_1, e_2, .., e_n\}$, другими словами определяется качественный состав вещества. Химическое вещество c определяется не только множеством обозначений химических элементов, но и количественным вхождением последних в состав вещества, раствора или смеси. Поэтому вещество c представлено кортежем (s, f), где $s \in S$, а f является отображением множества химических элементов, которые образуют вещество, на множество пар $R^+ \times R^+$, задающих соответственно минимальное и максимальное вхождение заданного химического элемента в вещество, раствор или смесь c. То есть $f : e_i \longrightarrow (R^+_{\min}, R^+_{\max})$, R^+ — множество неотрицательных действительных чисел. $R^+_{\min}$ и $R^+_{\max}$, соответственно, минимальная и максимальная концентрация химического элемента e_i в веществе c. В случае, когда концентрация конкретного химического элемента e_i в веществе c фиксирована, то $R^+_{\min} = R^+_{\max}$. Кристаллическая модификация m представляется кортежем (s, f, mod), где $s \in S$, $f : e_i \longrightarrow (R^+_{\min}, R^+_{\max})$, а mod — строковое обозначение модификации вещества.

Данные о свойствах химических сущностей, хранящиеся в разных ИС СНВМ с разной степенью детализации, должны быть представлены в виде строго типизированных наборов данных (возможно, реляционных таблиц). При этом связь сущностей из иерархии понятий со значениями свойств с помощью ER-диаграммы представима следующим образом (рис. 2.3.3).

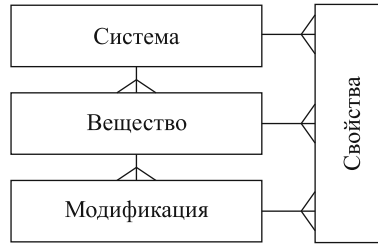


Рис. 2.3.3. Связь значений свойств химических объектов с иерархией понятий

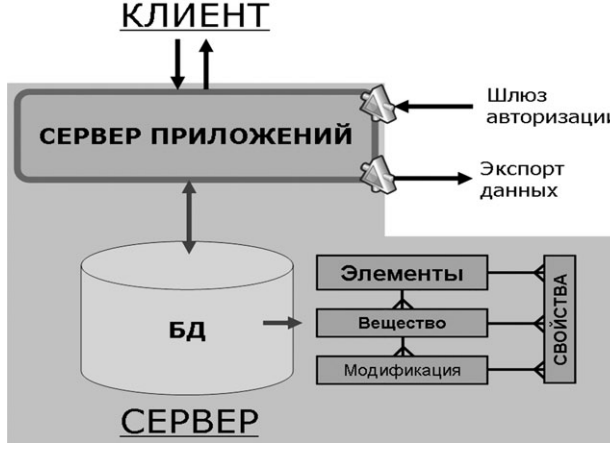


Рис. 2.3.4. Архитектура ИС СНВМ

Таким образом, предложенная схема данных должна быть реализована в ИС СНВМ для обеспечения поиска по качественному и количественному составу неорганических веществ с учетом кристаллических модификаций, а также для поиска по значениям конкретных физико-химических свойств.

На основании вышеизложенного, для построения современных ИС СНВМ с доступом через Интернет предлагается трехзвенная архитектура ИС СНВМ на основе систем хранения данных со строго типизированным учетом информации по: 1) качественному и количественному составу веществ; 2) кристаллическим модификациям; 3) значениям свойств. Следование этой архитектуре позволяет не только создавать ИС СНВМ с возможностью сложносоставных запросов, но и эффективно использовать их в качестве подсистем разработанной интегрированной ИС СНВМ (рис. 2.3.4).

2.4. Информационные системы по свойствам неорганических веществ ИМЕТ РАН

Как было отмечено в разделе 2.1, в настоящее время в мире существует множество ИС по свойствам веществ [4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 64–120]. При этом в каждой исследовательской организации создавались свои собственные информационные ресурсы, в которых накапливалась информация, относящаяся к ее тематике. Таким образом, в некоторых организациях исторически сформировались несколько центров хранения и обработки данных, что объясняется не только административными причинами, но и различиями в исследуемых предметных областях.

Примером организации, в которой существует несколько ИС, основанных на БД по свойствам неорганических веществ, может служить Институт металлургии и материаловедения им. А. А. Байкова Российской Академии Наук (ИМЕТ РАН). Эти ИС построены не только на различных программно-аппаратных платформах, но и с использованием разных подходов к хранению и обработке информации.

В разделе 2.3.2 были выявлены системные проблемы в информационных структурах, положенных в основу ряда зарубежных ИС СНВМ и предложены пути их решения с использованием обобщенной структуры данных ИС СНВМ. На основе предложенной в разделе 2.3.3 обобщенной структуры данных автором были созданы ИС «Кристалл» и ИС «Bandgap», которые обеспечивают требуемую гибкость при поиске данных для использования в компьютерном конструировании неорганических соединений.

2.4.1. Разработка ИС по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами «Кристалл»

В рамках исследований нами разработана ИС «Кристалл», построенная на основе БД Microsoft SQL Server 2008, содержащей экспериментальные данные о свойствах акустооптических, электрооптических и нелинейнооптических веществ, погрешностях, методах измерений, условиях получения и т. д. [69, 70].

Разработка базы данных ИС «Кристалл»

В основу разработки ИС «Кристалл» была положена разработанная совместно со специалистами-материаловедами концептуальная схема БД, представленная на рис. 2.4.1.

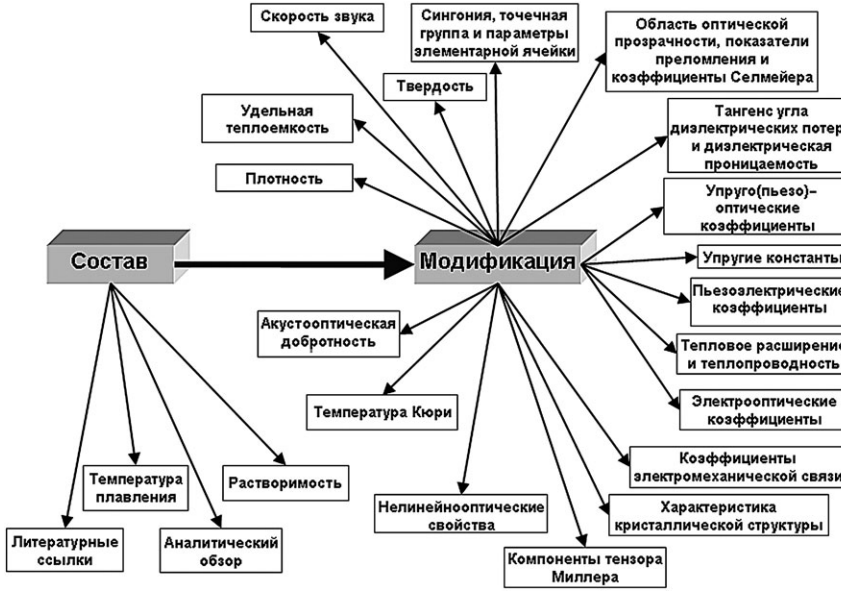


Рис. 2.4.1. Концептуальная схема базы данных ИС «Кристалл»

Как видно из представленной схемы, большинство свойств химических сущностей описывается на уровне кристаллической модификации. На основе концептуальной схемы была проведена нормализация БД и получена физическая структура данных для хранения информации в БД под управлением СУБД Microsoft SQL Server 2008, схематично показанная в табл. 2.2.

Все таблицы БД «Кристалл» (кроме Bibliogr) содержат общий столбец — номер соединения (вещества) HeadClue. Таблица HeadTabl — основная таблица БД, описывающая вещества. Таблицы БД условно разбиты на три группы:

- содержащие общие свойства соединения или общую для соединения информацию (общий столбец всех таблиц — номер соединения HeadClue) — HeadTabl, LitrTabl, SistTabl, SingTabl, HardTabl, DecrTabl, SuspTabl, AcopTabl, HeatTabl, DensTabl, PlavTabl, RefrTabl, CuryTabl, Wavepure;
- содержащие специфические для данной предметной области свойства, (общие столбцы всех таблиц — номер соединения HeadClue и обозначение сингонии SingCode) — ModfTabl, ElemTabl, MechTabl, Dielectr, Elastic1, ConstSel, RefrcInd, HeatExpn, ElOpTabl, NlOpTabl, MNopTabl, EsOpTabl, PzElTabl, DielDiss;

- служебные, содержащие названия графиков, поясняющих данные для конкретного соединения и свойства (общие столбцы для таблицы GrafTabl с другими таблицами — номер соединения HeadClue и с таблицей Property — номер свойства NompClue), содержащие номера свойств Property (общий столбец с таблицей GrafTabl — номер свойства — Nomprop = NompClue) и содержащие литературные ссылки Bibliogr (общий столбец с таблицами свойств — BkNumber).

Таблица 2.2. Схематическая структура данных в БД ИС «Кристалл»

HeadClue	**GrafTabl**	*NompClue*	
	Property	*Nomprop*	
HeadTabl			**Bibliogr**
HeadClue	**LitrTabl**		*BkNumber*
HeadClue	**HardTabl**		*BkNumber*
HeadClue	**DecrTabl**		*BkNumber*
HeadClue	**SuspTabl**		*BkNumber*
HeadClue	**AcopTabl**		*BkNumber*
HeadClue	**HeatTabl**		*BkNumber*
HeadClue	**DensTabl**		*BkNumber*
HeadClue	**PlavTabl**		*BkNumber*
HeadClue	**CuryTabl**		*BkNumber*
HeadClue	**Wavepure**		*BkNumber*
HeadClue	**SistTabl**		
		SingTabl	
		Singtype (*SingCode*)	
HeadClue	**ModfTabl**	*SingCode*	*BkNumber*
HeadClue	**ElemTabl**	*SingCode*	*BkNumber*
HeadClue	**MechTabl**	*SingCode*	*BkNumber*
HeadClue	**Dielectr**	*SingCode*	*BkNumber*
HeadClue	**Elastic1**	*SingCode*	*BkNumber*
HeadClue	**ConstSel**	*SingCode*	*BkNumber*
HeadClue	**RefrcInd**	*SingCode*	*BkNumber*
HeadClue	**HeatExpn**	*SingCode*	*BkNumber*
HeadClue	**ElOpTabl**	*SingCode*	*BkNumber*
HeadClue	**NlOpTabl**	*SingCode*	*BkNumber*
HeadClue	**MNopTabl**	*SingCode*	*BkNumber*
HeadClue	**EsOpTabl**	*SingCode*	*BkNumber*
HeadClue	**PzElTabl**	*SingCode*	*BkNumber*
HeadClue	**DielDiss**	*SingCode*	*BkNumber*

Краткое описание всех реляционных отношений БД «Кристалл» приведено в табл. 2.3.

Таблица 2.3. Реляционные таблицы БД «Кристалл»

Название	Назначение таблицы
HeadTabl	ключевая таблица БД, содержащая информацию о количественном составе вещества и данные по специалистам, проводившим экспертную оценку
GrafTabl	данные по графической информации
Properties	перечень свойств, информация о которых хранится в ИС
Bibliogr	справочник литературных ссылкок
SistTabl	количественный состав веществ
HardTabl	твердость
SuspTabl	растворимость (разных растворителях)
DensTabl	плотность
PlavTabl	температуры плавления
CuryTabl	температура Кюри
HeatTabl	теплоемкость
AcopTabl	акустооптические свойства
DecrTabl	характеристики распространения и затухания упругих волн
Wavepure	области прозрачности кристаллов
LitrTabl	привязка литературных ссылок из Bibliogr к веществам из HeadTabl
SingTabl	сингонии кристаллических решеток различных полиморфных модификаций веществ
ModfTabl	симметрия и условия существования различных полиморфных модификаций веществ
ElemTabl	параметры элементарной ячейки
HeatExpn	тепловое расширение и теплопроводность
Dielectr	диэлектрические постоянные
DielDiss	диэлектрические потери
PzElTabl	пьезоэлектрические коэффициенты
MechTabl	коэффициенты электромеханической связи
Elastic1	упругие постоянные
RefrcInd	показатели преломления
ConstSel	коэффициенты Селмейера
NlOpTabl	нелинейнооптические коэффициенты
MnOpTabl	компоненты тензора Миллера
ElOpTabl	электрооптические коэффициенты
EsOpTabl	упругооптические коэффициенты

Каждая таблица создавалась с использованием SQL DDL-операторов. Например, основная таблица HeadTabl, содержащая список соединений, создавалась с помощью SQL-скрипта:

```
CREATE TABLE [dbo].[HeadTabl](
        [HeadClue] [int] NOT NULL,
        [System] [varchar](128) NOT NULL,
        [Expert] [varchar](32) NOT NULL,
        [Help] [varchar](32) NOT NULL,
        [Class] [int] NOT NULL,
CONSTRAINT [PK_HeadTabl] PRIMARY KEY CLUSTERED
( [HeadClue] ASC) ON [PRIMARY]
) ON [PRIMARY];
```

В итоге получаем реляционную структуру данных, лишь малая часть которой отображена на рис. 2.4.2 с указанием назначения таблиц. Важно отметить, что информационное наполнение БД ИС «Кристалл» осуществлялось с помощью программного комплекса DBAdmin, разработанного автором (см. описание в разделе 2.4.6).

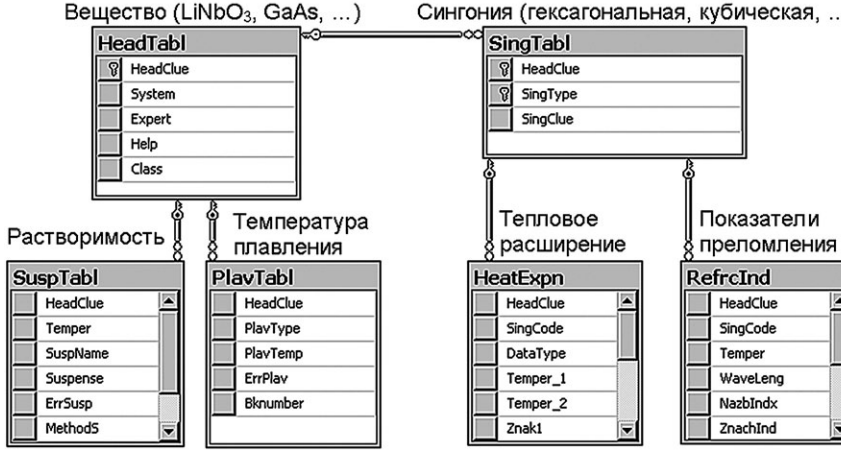


Рис. 2.4.2. Реляционная модель части БД ИС «Кристалл»

Разработка Web-интерфейса ИС «Кристалл»

Программное обеспечение ИС «Кристалл» является Web-интерфейсом к серверу баз данных Microsoft SQL Server 2008. Для организации доступа к БД использована технология Active Server Pages (ASP) с поддержкой ActiveX Data Objects (ADO). В качестве Web-сервера использован Microsoft Internet Information Server 7.0, который наиболее тесно интегри-

рован с используемой операционной системой MS Windows 2008 Server. На Web-сервере хранятся ASP-документы (шаблоны страниц), с помощью которых запросы пользователей пересылаются серверу БД. Данные от сервера БД обрабатываются Web-приложением и оформляются в виде HTML-страниц с использованием каскадных таблиц стилей (CSS), отсылаемых пользователям. Таким образом, конечный пользователь получает доступ к информации БД с помощью любой программы-браузера (требуется поддержка JavaScript и Cookies).

Разработанная программная подсистема обработки типовых запросов выполняет следующие функции:

- реализацию санкционированного доступа к БД «Кристалл» для надежного функционирования в условиях сети Интернет;
- формирование SQL-запросов к БД «Кристалл» и вывод результатов поиска;
- организацию удобного пользовательского интерфейса.

Главное меню обработки запросов БД «Кристалл» позволяет пользователю выбрать режим работы: просмотр данных о заданном веществе или возможность создания сложного запроса. Работая в режиме «сложный запрос», пользователь может выбрать соединения, имеющие заданные свойства или совокупность свойств, что важно при формировании обучающей выборки для компьютерного конструирования неорганических соединений. При работе в этом режиме пользователь сначала должен выбрать необходимые ему свойства (рис. 2.4.3).

Затем задать в соответствующих формах-таблицах диапазоны значений для поиска веществ (рис. 2.4.4). Пользователь должен ввести в поля форм значения для поиска (если какое-то поле остается пустым, то оно не будет учитываться при поиске в БД). Например, пользователь может подобрать вещества, значения твердости которых по Моосу превышают 1.2 ГПа, причем коэффициент их термического расширения лежит в пределах $(0.05–1)\cdot10^{-6}$ $К^{-1}$, а коэффициент линейного электрооптического r_{ijk} эффекта выше $20\cdot10^{-12}$ $м/В^2$. Этот запрос соответствует реальным требованиям к электрооптическому кристаллу. Например, общим требованием для кристаллов этого типа является достаточно высокая твердость, обеспечивающая хорошую обрабатываемость поверхностей рабочих элементов и лучшую сохранность их в процессе эксплуатации (твердость должна быть выше 1.2 ГПа).

Пример выдачи списка соединений, параметры которых удовлетворяют сформулированному выше запросу показан на рис. 2.4.5.

Далее пользователь может посмотреть информацию об отобранном веществе, выбрав нужное соединение. Данные выдаются в табличной и графической формах, Возможен просмотр соответствующих литературных ссылок и полных текстов статей.

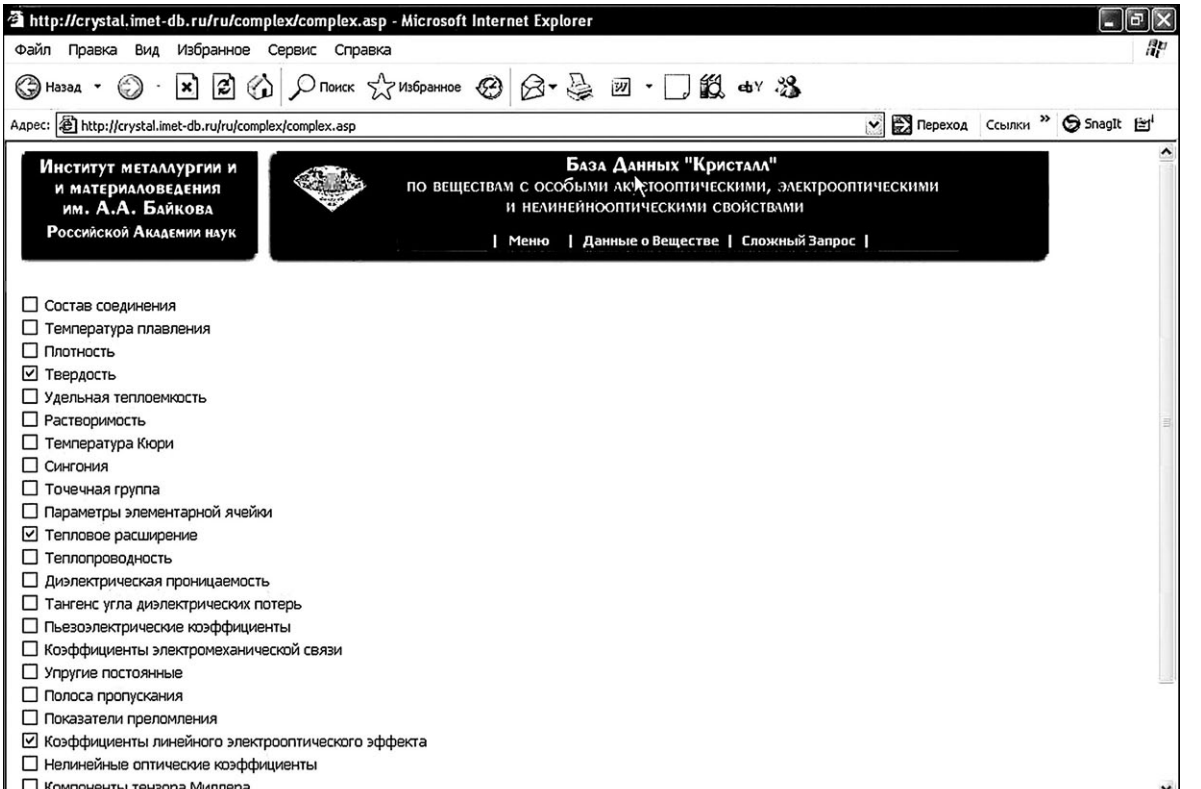


Рис. 2.4.3. Форма выбора свойств в сложном запросе ИС «Кристалл»

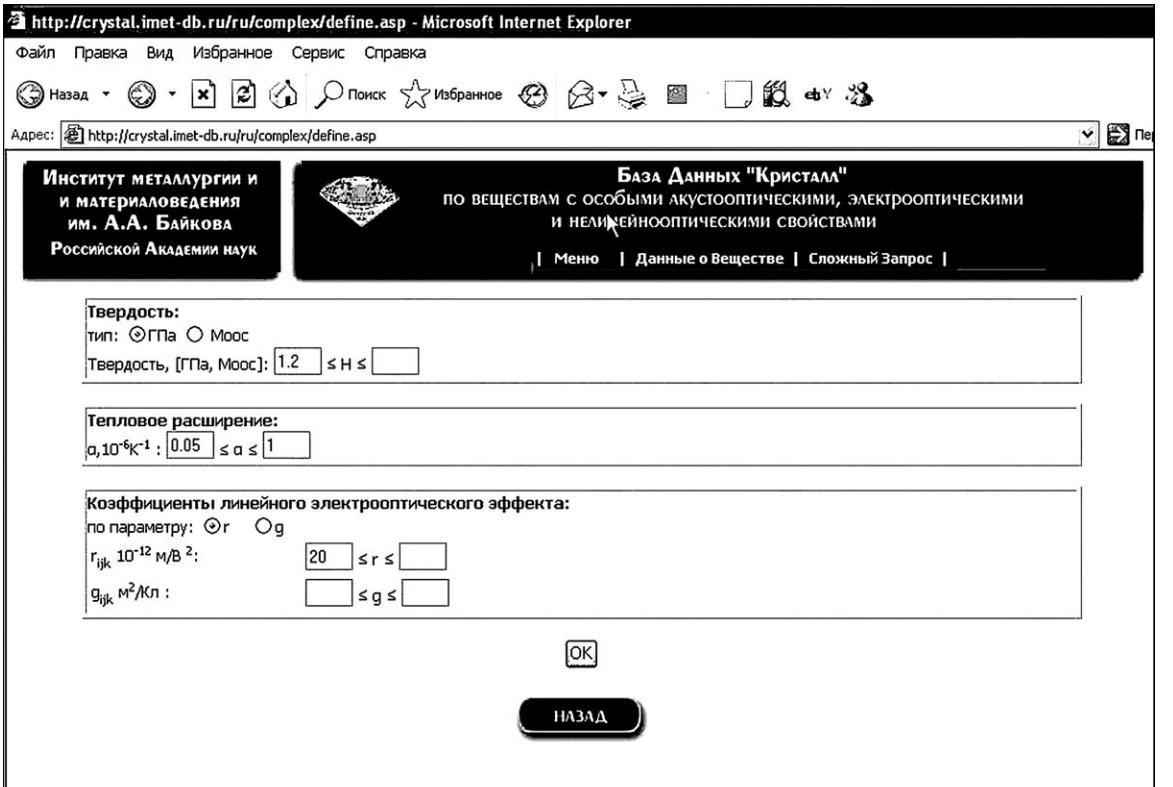


Рис. 2.4.4. Форма для ввода ограничений на значения свойств в сложном запросе к ИС «Кристалл»

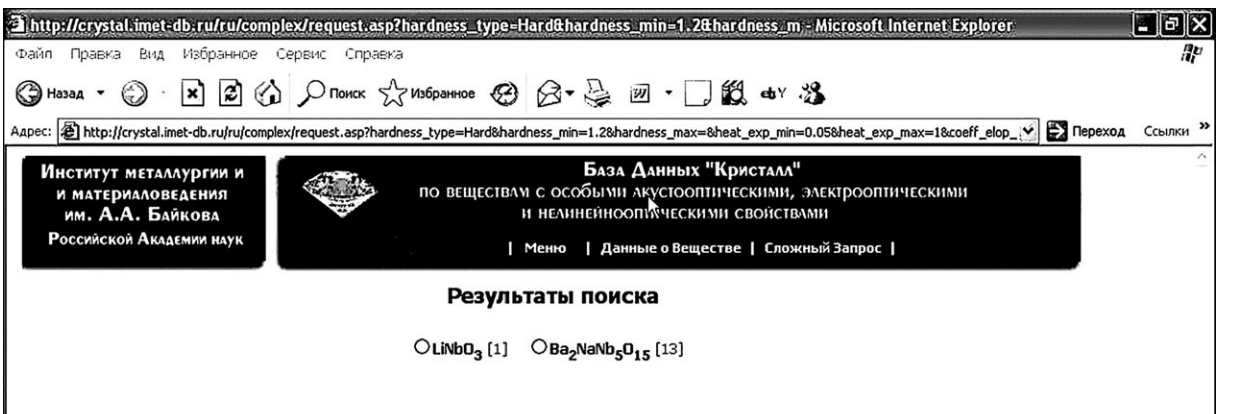


Рис. 2.4.5. Результаты поиска при выполнении сложного запроса в ИС «Кристалл»

Существенно, что основная часть информации БД собрана и оценена российскими специалистами, непосредственно участвующими в разработке и использовании химических соединений, относящихся к этим классам веществ. Информация по свойствам каждого вещества дополнена аналитическим обзором, в котором кратко описана технология получения веществ, возможные области их применения. В состав аналитического обзора включены также данные об особых свойствах веществ, которые не было возможно отобразить в рамках жесткой информационной структуры реляционной БД. Там же, по возможности, дана экспертная оценка данных, хранящихся в БД.

В настоящее время ИС содержит следующую информацию более чем о двух сотнях наиболее важных для практических применений веществ: данные о составе соединения, температуре плавления, плотности, твердости, удельной теплоемкости, растворимости, температуре Кюри, кристаллической структуре и параметрах элементарной ячейки, тепловом расширении, теплопроводности, диэлектрической проницаемости, диэлектрических потерях, пьезоэлектрических свойствах, коэффициентах электромеханической связи, упругих постоянных, полосе пропускания и показателях преломления кристалла, коэффициентах Селмейера, электрооптических свойствах, нелинейных оптических свойствах, пьезо(упруго-)оптических коэффициентах, скоростях распространения упругих волн и коэффициентах затухания, акустооптических свойствах при различных условиях. ИС содержит обширную графическую информацию о зависимостях свойств (более 1000 рисунков). Для большинства англоязычных публикаций последних лет возможен просмотр полных текстов статей, из которых извлечена информация, хранящаяся в таблицах БД.

Стоит отметить, что в настоящее время разработаны русскоязычная и англоязычная версии ИС «Кристалл». Информация в БД ИС обновляется с использованием специализированного ПО [136, 139]. Для зарегистрированных пользователей ИС «Кристалл» доступна из глобальной сети Интернет (http://crystal.imet-db.ru). Также доступ к информации, отфильтрованной по выбранному веществу, все желающие могут получить из единой точки входа, рассматриваемой в главе 6 настоящей работы (http://meta.imet-db.ru).

2.4.2. Разработка ИС по ширине запрещенной зоны неорганических соединений «Bandgap»

В рамках исследований разработана ИС «Bandgap», которая содержит информацию по ширине запрещенной зоны (*Eg*) основных неорганических соединений. Разработка данной ИС является продолжением

исследований по сбору и систематизации данных по свойствам веществ и материалов для электронной промышленности, т. к. ширина запрещенной зоны является одной и важнейших характеристик полупроводниковых веществ. По величине *Eg* можно судить о типе химической связи, доминирующей в соединении, устойчивости соединения в определенном интервале изменений состава и внешних параметров, типе электронной проводимости в образцах, склонности вещества к ионной проводимости, а также основных термодинамических характеристиках соединения (энтальпии образования $\Delta H_{f\,298}$, энтропии ΔSm и температуры плавления *Tm* и других) [141, 142].

ИС «Bandgap» содержит значения *Eg*, полученные из оптических, термических и электрофизических измерений. Приведены сопутствующие данные по химическому и фазовому составу веществ, типу кристаллической решетки (система, структурный тип, пространственная группа), указана ориентации образцов и температура измерений. Все приведенные данные соответствуют нормальному атмосферному давлению, а также отсутствию заметных электрических и магнитных полей.

Архитектурные особенности построения ИС «Bandgap» во многом повторяют успешный опыт, полученный при разработке ИС «Кристалл», описанной ранее. Поэтому кратко рассмотрим основные результаты полученные в процессе создания ИС «Bandgap». Основу ИС составляет реляционная БД под управлением Microsoft SQL Server 2008, содержащая описания неорганических соединений, численные данные по ширине запрещенных зон неорганических соединений и ссылки на библиографические источники и дополнительные графические материалы (рис. 2.4.6).

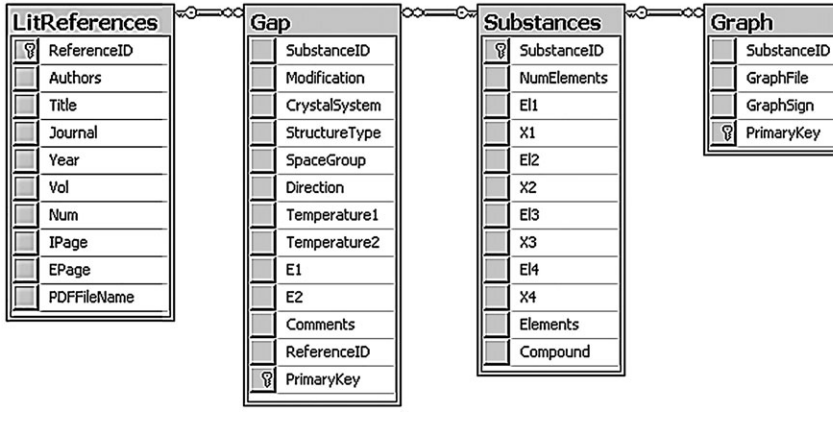


Рис. 2.4.6. Логическая схема данных ИС «Bandgap» в Microsoft SQL Server 2008

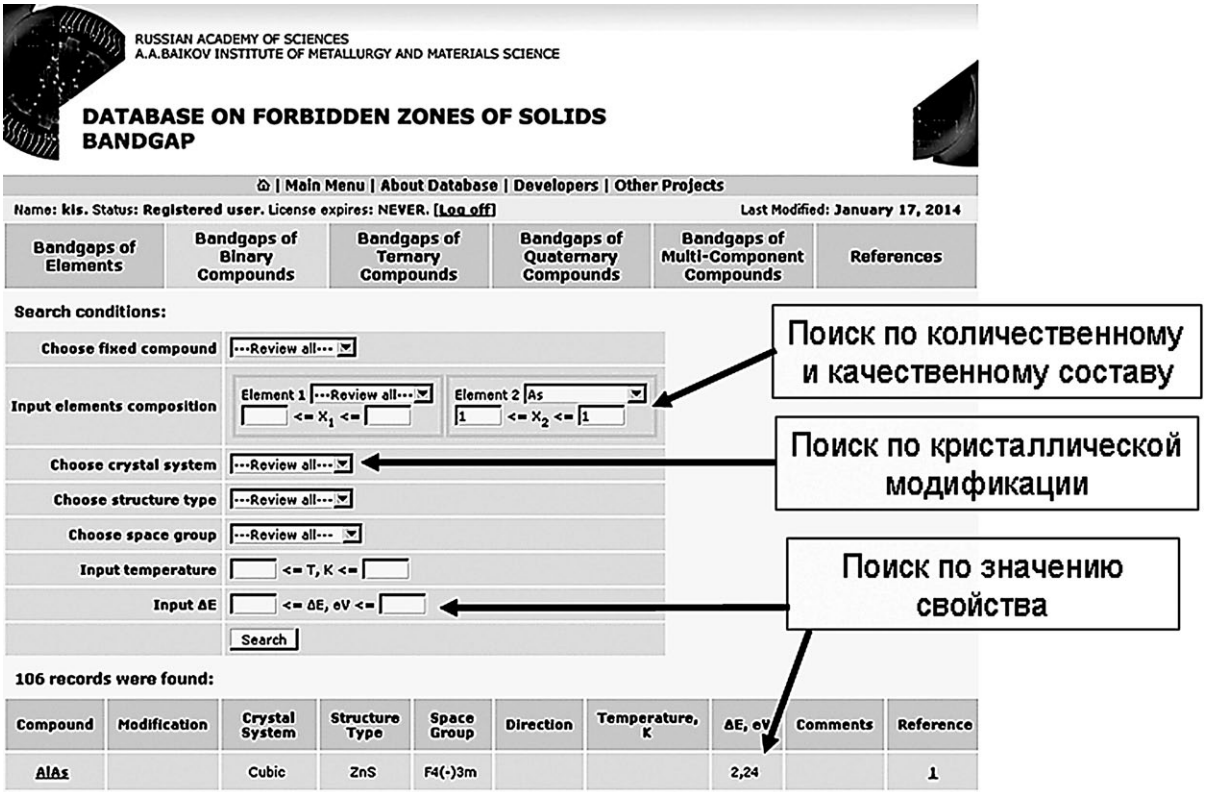


Рис. 2.4.7. Пример поискового запроса в ИС «Bandgap»

ИС была разработана в виде Web-приложения, реализованного с использованием технологии Active Server Pages (ASP). Для доступа к данным, хранимым на Microsoft SQL Server 2008, используется интерфейс ActiveX Data Objects (ADO). Web-приложение ИС развернуто на Web-сервере Microsoft IIS 7.0 под управлением Microsoft Windows 2008 Server в ИМЕТ РАН.

Пользовательский интерфейс является достаточно удобным для поиска данных сразу по количественному и качественному составу вещества, типу кристаллической структуры и диапазону значений ширины запрещенной зоны, что полезно при составлении обучающих выборок для прогнозирования ширины запрещенной зоны методами компьютерного конструирования неорганических соединений (рис. 2.4.7).

Язык ИС — английский, что обеспечивает возможность работы с ней не только для российских пользователей, но и для зарубежных специалистов. До недавнего времени ИС «Bandgap» была доступна из глобальной сети Интернет только для зарегистрированных в ИМЕТ РАН пользователей, сейчас свободный доступ к этой ИС открыт для всех желающих по адресу http://bg.imet-db.ru (внимание, необходима online-регистрация). Также доступ к информации этой ИС можно получить из единой точки входа, рассматриваемой в главе 6 настоящей работы (http://meta.imet-db.ru).

2.4.3. ИС по свойствам неорганических соединений «Фазы»

ИС по свойствам неорганических соединений «Фазы» [1] предназначена для хранения информации о неорганических соединениях. В настоящее время в БД ИС содержатся сведения более чем о 43 тыс. тройных соединений из 18 тыс. тройных систем. Информация в БД ИС обновляется ежедневно. Часть информации оценена экспертами. Подготовлен к вводу массив данных по свойствам четверных соединений, содержащий информацию более чем о 50 тыс. соединений. ИС содержит сведения о наиболее распространенных и изученных характеристиках соединений. Более детальная информация хранится в специализированных ИС: «Диаграмма», «Кристалл» и т. д. (см. ниже).

Информация для формирования БД ИС «Фазы» отбирается из периодических изданий, справочников, монографий, отчетов, а также реферативных журналов. Анализ литературы показывает, что наиболее часто упоминаемая характеристика неорганических соединений — сингония. Следующей по степени изученности является пространственная группа и число формульных единиц в элементарной ячейке. Наиболее распространенная (по степени изученности) термохимическая характеристика — тип и температура плавления, далее — температура распада тройного соеди-

нения в твердой или газообразной фазах. Значительно реже определяется температура кипения. Информация обо всех этих свойствах неорганических соединений включена в БД ИС «Фазы».

Пользователями ИС «Фазы» являются специалисты по неорганической химии. ИС доступна зарегистрированным пользователям из глобальной сети Интернет (http://phase.imet-db.ru).

2.4.4. ИС по фазовым диаграммам систем с полупроводниковыми фазами «Диаграмма»

Основной информацией ИС «Диаграмма» [65, 66, 67, 71, 72] являются: хранимые в БД таблицы собранных и оцененных экспертами экспериментальных данных по линиям многовариантных, моновариантных и нонвариантных равновесий; данные по особым точкам; таблицы данных по указанным выше линиям равновесий и особым точкам, полученные в результате статистической обработки или термодинамического согласования экспериментальных данных; рисунки фазовых диаграмм. Помимо этого в БД ИС хранятся сведения о кристаллической структуре фаз, файлы термодинамических свойств фаз и систем, файлы расчетных моделей, а также написанные экспертами аналитические обзоры по фазовым диаграммам, в которых, в частности, дается дополнительная информация по системам, не включенная в таблицы БД. Структура БД подробно описана в публикации, посвященной данной ИС [65]. В настоящее время БД содержит информацию о нескольких десятках двойных и тройных систем, извлеченную из почти 2 тыс. публикаций. Содержимое БД постоянно пополняется новыми данными [136, 137].

При разработке БД «Диаграмма» особое внимание было уделено оценке достоверности хранящейся информации о фазовых диаграммах. К сбору и оценке качества данных были привлечены специалисты РАН, НИИ и вузов, имеющие опыт исследования полупроводниковых систем. Достоверность измерения каждого экспериментального значения (содержания компонентов, температуры, давления и т. д.) в таблицах БД оценивалось экспертами по пятибалльной шкале, соответствующей различным фиксированным уровням ошибок измерения. Информация о величине ошибки, предлагаемая экспертом, выдается пользователю при просмотре соответствующих таблиц, содержащих информацию о линиях равновесий.

Рисунки фазовых диаграмм можно просмотреть в статическом и динамическом режимах. В первом случае пользователь может визуализировать на экране, напечатать и записать в файл рисунок в формате jpeg [138]. Динамический режим требует дополнительной установки на компьютере

пользователя компонента Web-браузера — Macromedia Flash (современные версии всех без исключения браузеров поддерживают эту технологию) и дает возможность динамического определения координат точек на T-x фазовых диаграммах и масштабирования рисунков, что позволяет просмотреть и распечатать наиболее интересные для пользователя области диаграмм.

Дополнительно к табличной и графической информации ИС «Диаграмма» предоставляет возможность доступа к полным текстам большинства англоязычных статей последних лет, ссылки на которые указаны в аналитических обзорах для соответствующих систем. Для зарегистрированных пользователей ИС «Диаграмма» доступна из глобальной сети Интернет (http://diag.imet-db.ru).

2.4.5. ИС по свойствам кремния «Кремний»

ИС «Кремний» создавалась на основе информации, собранной специалистами Гиредмет, ИХПМ и ИМЕТ РАН с 1985 г. Эта специализированная ИС содержит сведения исключительно по свойствам полупроводникового кремния. Значимость информации в этой ИС трудно переоценить, т. к. именно кремний является основным полупроводниковым материалом, на долю которого приходится более 90 % общего мирового объема производства. В перспективе лидирующая роль кремния в приборостроении сохранится. ИС содержит информацию о кремнии, начиная с сырья; а именно, о процессах получения, очистке, выращивании монокристаллов, пластинах кремния, эпитаксиальных структурах, методах контроля свойств. Также ИС содержит информацию о физико-химических свойствах кремния, о его мировых производителях и потребителях, о мировом уровне промышленного производства кремния, о конкурентоспособных показателях, требуемых для выхода кремниевой продукции на мировой рынок, о тенденции развития, об аппаратуре, используемой в современной технологии и о направлениях ее развития и т. д. Для зарегистрированных пользователей ИС «Кремний» доступна из глобальной сети Интернет (http://si.imet-db.ru).

2.4.6. Разработка программного комплекса для удаленного администрирования гетерогенных БД ИМЕТ РАН

Организация удаленного администрирования и эффективной поддержки баз данных посредством сетей, относится к одной из важнейших задач информационных технологий. Это обеспечивает оперативность

информационного наполнения баз данных. Как было показано выше, в ИМЕТ РАН существует целый ряд БД по различным физико-химическим свойствам веществ, которые имеют различную структуру и функционируют под управлением различных СУБД, таких как Microsoft SQL Server, Oracle и Postgres на платформах Microsoft Windows и Sun Solaris. Для всех этих баз данных существуют свои программы редактирования информации, которые были разработаны разными программистами и имеют сильно отличающиеся пользовательские интерфейсы. Число такого рода программ растет с увеличением числа баз данных. Все это в значительной степени затрудняет редактирование информации экспертами ИМЕТ РАН.

Поэтому возникла необходимость создания единого универсального механизма удаленного администрирования различных баз данных ИМЕТ РАН, обеспечивающего полную функциональность, необходимую для полноценной работы с любой из существующих БД. Для обеспечения универсального взаимодействия с различными СУБД необходимо было выбрать мощный и гибкий механизм работы с удаленными источниками данных. В качестве альтернатив рассматривались следующие интерфейсы доступа к БД с платформы Windows:

- **ODBC** (Open Database Connectivity) — низкоуровневый интерфейс доступа к реляционным базам данных;
- **MFC** (Microsoft Foundation Classes) **ODBC** classes — высокоуровневый интерфейс доступа к реляционным базам данных, основанный на ODBC;
- **DAO** (Data Access Objects) — высокоуровневый интерфейс доступа к реляционным базам данных, основанный на использовании Access/Jet процессора баз данных;
- **RDO** (Remote Data Objects) — высокоуровневый интерфейс доступа к реляционным базам данных, созданный для программистов на Visual Basic;
- **OLE DB** (Object-Linking and Embedding Database) — новый низкоуровневый интерфейс доступа к источникам данных от Microsoft;
- **ADO** (ActiveX Data Objects) — новый высокоуровневый, основанный на OLE DB, интерфейс доступа к источникам данных;
- **ADO.Net** (ActiveX Data Objects .Net) — новейший высокоуровневый интерфейс доступа к источникам данных.

Для выбора интерфейса доступа к БД использовались следующие критерии:

- **Наличие объектной модели** — обеспечивает ли интерфейс объектную модель, которая облегчает написание объектно-ориентированных программ.

- **Поддержка нереляционных источников данных** — обеспечивает ли интерфейс доступ к данным, хранимым в нереляционных источниках данных (т. к. все интерфейсы позволяют осуществлять доступ к реляционным источникам данных, критерий **поддержка реляционных источников данных** исключен).
- **Возможность низкоуровневого контроля** — обеспечивает ли интерфейс возможность низкоуровневого доступа к серверам реляционных баз данных.
- **Высокая производительность** — способен ли интерфейс обеспечить высокую производительность при взаимодействии с СУБД.
- **Поддержка разъединенной модели работы** — обеспечивает ли интерфейс возможность использования разъединенной парадигмы при работе с источником данных.
- **Соотношение «функциональность / объем кода»** — показывает, как много кода нужно написать по сравнению с функциональностью, которую можно получить от него.

В табл. 2.4 содержатся результаты сравнительного анализа сильных и слабых сторон различных интерфейсов доступа к базам данных. Приняты следующие условные обозначения: знак «+» обозначает сильную сторону, «+ +» обозначает особенно сильную сторону, «–» обозначает слабую сторону, и наконец, отсутствие знаков означает отсутствие значимых достоинств и недостатков.

Таблица 2.4. Сравнение интерфейсов доступа к базам данных

интерфейс / критерий	ODBC	MFC ODBC	DAO	RDO	OLE DB	ADO	ADO.Net
Объектная модель	–	+	+	+	+	+ +	+ +
Нереляционные источники данных	–	–	–	–	+	+	+
Низкоуровневый контроль	+		–	–	+		
Производительность	+		–		++	+	+
Разъединенная модель работы с БД	–	–	–	–	–	–	+
Соотношение «функциональность / объем кода»	–		+		–	+	+

Из всех рассмотренных интерфейсов только OLE DB, ADO и ADO.Net подходят для использования в программном комплексе. OLE DB является мощным, но при этом низкоуровневым интерфейсом. Это потребует больше кода и затрат при реализации, чем потребовали бы высокоуровневые интерфейсы, такие как ADO и ADO.Net. Учитывая то, что функциональные возможности ADO и ADO.Net достаточны для реализации программного комплекса, и низкоуровневого контроля, предлагаемого OLE DB, не требуется, интерфейс OLE DB был исключен из возможных интерфейсов для реализации программного комплекса.

Интерфейс ADO предоставляет простую и гибкую объектную модель, обладающую при этом хорошей производительностью. Это делает ADO наилучшим решением для разработки клиент-серверных приложений для баз данных, для которых необходимо постоянное подключение к источнику данных, как и в случае программ удаленного администрирования. Интерфейс ADO.Net является наиболее пригодным для разработки взаимодействующих с БД Web-приложений, обеспечивая меньшую загрузку серверов БД. Учитывая то, что программа удаленного администрирования реляционных баз данных должна поддерживать постоянное соединение с сервером БД для динамического отслеживания изменений данных и в структуре БД и являться Win32-приложением оптимальным выбором для доступа к данным в нашем случае является интерфейс ADO.

С использованием программного интерфейса ADO был разработан универсальный программный комплекс DBAdmin, позволяющий выполнять удаленное администрирование БД всех ИС ИМЕТ РАН с использованием единого пользовательского интерфейса (рис. 2.4.8). Комплекс разработан на языке C++ с использованием RAD-среды Borland C++Builder 6 Enterprise Edition [137]. Особенностями данного комплекса являются:

- возможность удаленного (по локальной сети или через Интернет) взаимодействия с БД;
- способность работать с БД произвольной структуры, поскольку структура данных считывается из каталога при подключении к информационному источнику;
- возможность эффективно взаимодействовать с БД под управлением разных СУБД (Microsoft SQL Server, Oracle) за счет использования стратегии Microsoft Universal Data Access и технологий OLE DB и ODBC.

Использование DBAdmin позволяет стандартизировать процедуры администрирования всех БД в рамках ИМЕТ РАН и дает возможность использования это программного комплекса для единого управления всеми БД в рамках интегрированной ИС.

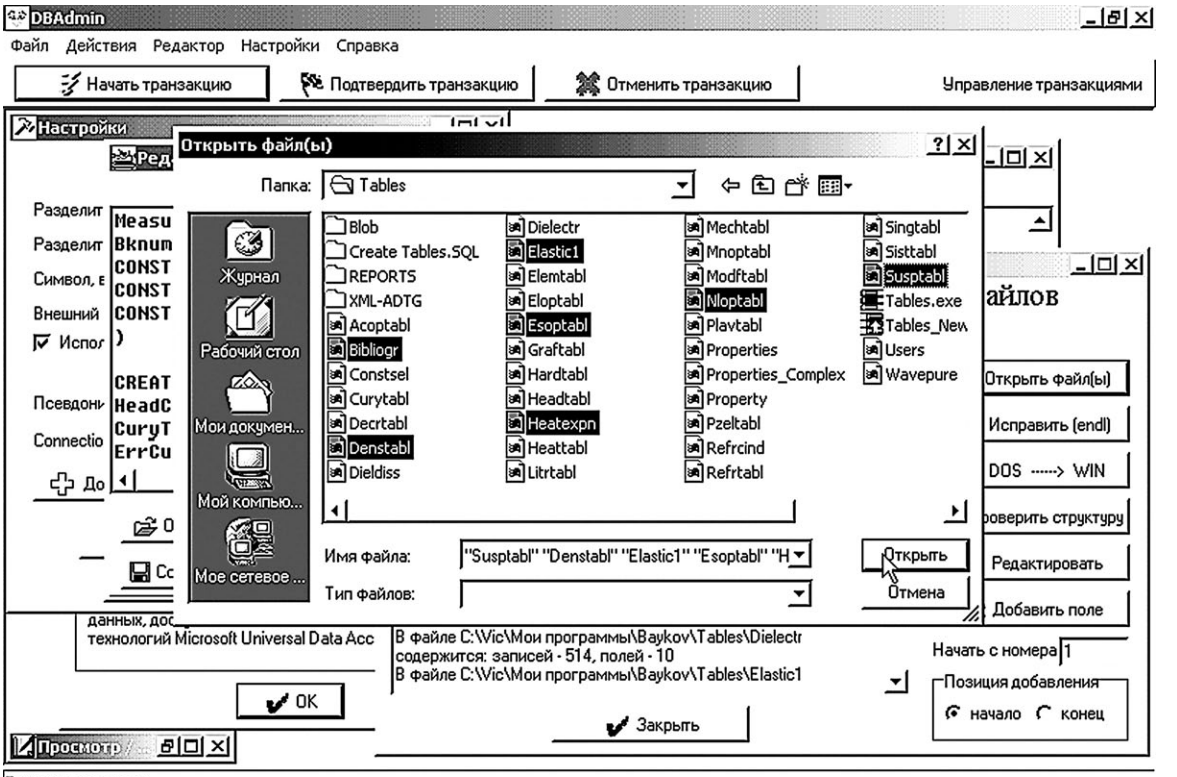


Рис. 2.4.8. Интерфейс программного комплекса DBAdmin

2.4.7. Особенности ИС ИМЕТ РАН

Информационные системы ИМЕТ РАН были разработаны в разное время, что отражается не только на их реализации на различных программно-аппаратных платформах, но и в использовании разных подходов к хранению и обработке информации. ИС ИМЕТ РАН являются репрезентативными для проведения анализа их архитектуры с целью разработки общей схемы интеграции ИС в данной предметной области. Следует отметить, что важными особенностями рассмотренных ИС по свойствам неорганических веществ, разработанных в ИМЕТ РАН, являются:

- достоверность информации, обеспеченная тем, что сбор и экспертную оценку данных выполняют высококвалифицированные специалисты;
- постоянное обновление данных;
- простота использования созданных ИС за счет тщательной проработки программного обеспечения и создания удобных для пользователей и операторов интерфейсов;
- возможность оперативного доступа к информации из сети Интернет, обеспеченная высокоскоростной волоконно-оптической линией связи и мощным аппаратным обеспечением ИМЕТ РАН (сервера БД и Web-сервера).

2.5. Расчетные подсистемы информационных систем по свойствам неорганических веществ

Очевидно, что все многообразие информации, доступ к которой осуществляется из специализированных информационных систем, невозможно представить посредством использования только табличной формы представления данных. Поэтому исследователи, работающие над изучением физико-химических свойств того или иного вещества, активно используют графические материалы (рисунки, графики зависимостей и т. п.). Нередки также случаи написания расчетных подпрограмм, с помощью которых динамически рассчитываются значения тех или иных свойств заданного класса веществ по введенным пользователем параметрам или осуществляется визуализация рассчитанной по некоторым правилам информации.

Существенной особенностью современных ИС по свойствам неорганических веществ является необходимость хранения и предоставления пользователям не только информации в табличной форме, но и множества графической информации и расчетных подсистем.

Например, в ИС «Диаграмма» необходимой задачей являлось отображение рисунков фазовых диаграмм (ФД) для конечного пользователя ИС. Отметим, что возможны различные методы представления рисунков ФД [143]:

- использование термодинамических моделей элементов ФД с возможностью последующей визуализации рассчитанных кривых, что влечет за собой написание расчетно-визуализирующей подсистемы [144, 145, 146];
- использование различных аппроксимирующих математических моделей элементов ФД (обычно полиномов) также с последующей визуализацией рассчитанных кривых (написание расчетно-визуализирующей подсистемы) [144, 147, 148, 149];
- табличное представление данных и использование интерполирующих сплайнов для расчета данных между узлами таблицы (координаты точек из таблицы используются расчетно-визуализирующей подсистемой в качестве входных данных для визуализации ФД с помощью построения сплайнов) [149, 150, 151];
- графический метод, при котором рисунок вводится в ЭВМ с помощью какого-либо специального устройства (дигитайзера, сканера и т. п.) и преобразуется в формат, удобный для хранения в вычислительной машине; обратное преобразование переводит данные в графическую форму (хранение ФД в виде изображения и его последующая визуализация для конечного пользователя ИС) [145, 148].

Учитывая недостатки использования термодинамических моделей (небольшое количество систем с надежно определенными термодинамическими свойствами фаз, неточность термодинамических моделей фаз, которые позволяют рассчитать фазовые диаграммы со значительно большей погрешностью, чем это дает экспериментальное построение, громоздкость термодинамических расчетов, что значительно замедляет ответ даже на очень простые запросы), подробно проанализированные в [143, 150], было принято решение отказаться от их использования.

Использование различных математических моделей, а также интерполирующих сплайнов для аппроксимации элементов ФД также имеет множество недостатков. Сложность выбора вида аппроксимирующих полиномов и описания кривых фазовых равновесий вблизи сингулярных точек делает применение этого подхода крайне сложным.

Для визуализации рисунков фазовых диаграмм был выбран графический метод, который является наиболее технологичным из всех рассмотренных. Его недостатком является только невозможность обработки данных в случае многомерных пространственных фигур. Однако, как правило, в этом случае для представления данных о фазовых диаграммах используются проекции или сечения многомерной фигуры.

Фазовые диаграммы в ИС «Диаграмма» хранятся в виде растровых и векторных изображений. Рисунок в растровом формате представляет собой матрицу конечного размера из точек, цвет которых точно определен. Растровое изображение высокого разрешения обеспечивает требуемую точность вывода деталей ФД. Использован сжатый формат хранения растровой графики (jpeg, gif) [138], что позволяет в значительной степени сократить объем файла хранимого изображения и, тем самым, ускорить процесс передачи информации в сети Интернет. Однако, в связи с тем, что в этом случае рисунок представлен в виде набора пикселей, программное масштабирование растрового изображения бессмысленно, вместо этого необходимо получить изображение более высокого качества, объем которого при двукратном увеличении будет в четыре раза больше исходного. Следует также учесть, что привязка к системе координат рисунка при использовании растровой графики не имеет смысла ввиду отсутствия возможности масштабирования.

В связи с этим дополнительно в ИС хранятся рисунки в векторном формате Macromedia Shockwave Flash (swf) [152]. Рисунок в векторном формате представляет собой множество примитивных графических объектов, таких, как линия, кривая, область заливки, текст (символ) и др. В файле хранятся математические описания этих объектов, а проекция такого рисунка на растровое устройство (каковым является монитор компьютера) является сложным, многоэтапным процессом. Этим занимается особая программа (RIP — Raster Image Processor). Вначале RIP выстраивает структуру рисунка, распределяя объекты по уровням, строит пересечения и объединения объектов. В связи с тем, что описания векторных графических объектов являются аналитическими, возможно масштабирование таких рисунков в любых пропорциях без потери качества изображения. Используемый формат swf является векторным графическим форматом. Он широко распространен и поддерживается большинством современных браузеров. Программы установки браузеров зачастую предлагают при установке включить в состав устанавливаемого программного обеспечения бесплатную систему для просмотра графических файлов этого типа — Macromedia Flash Viewer. В случае отказа клиента от установки проектора вместе с браузером (например, в целях экономии места на диске, хотя проектор занимает всего порядка 350 килобайт) установка может быть произ-

ведена при первом обращении к файлу данного типа, что потребует от пользователя минимальных усилий — согласиться на предложение браузера установить компоненту. При этом загрузка и установка программной компоненты будут произведены автоматически.

Графическая информация хранится в файле в сжатом виде, что позволяет существенно сократить объем передаваемого в сети Интернет файла. В среде Macromedia Flash легко решаема проблема распечатки изображения на принтере пользователя, т. к. такая функция встроена в проектор. Для разметки координатной сетки на рисунках фазовых диаграмм разработана специальная система D_Mapper. Она позволяет осуществить привязку графиков к системе координат. Таким образом, пользователи ИС, перемещая курсор мыши по графику, имеют возможность отслеживать текущие координаты курсора в заданной системе координат.

Отображение графической информации на компьютере пользователя осуществляется обычными средствами программы просмотра (браузера). На рис. 2.5.1 приведен пример выдачи графической информации из ИС «Диаграмма».

Существуют и другие типы расчетных подсистем, активно использующиеся в информационных системах по свойствам веществ. Например, в ИС по процессам получения эпитаксиальных гетероструктур полупроводниковых материалов методом жидкофазной эпитаксии [113], разработанной в МИТХТ им. М. В. Ломоносова, используется информационно-расчетная подсистема для компьютерного моделирования процессов жидкофазной эпитаксии. Особенность данной расчетной подсистемы, как и большинства расчетных подсистем, заключается в использовании информации, хранимой в БД указанной ИС. То есть расчетная система жестко «привязана» к информационным структурам, содержащимся в БД, и не может быть использована отдельно от указанной БД, например, самостоятельно или в рамках другой информационной системы.

Отметим, что основная особенность сложных систем визуализации информации, как и расчетных подсистем, заключается в невозможности их безболезненного использования вне рамок информационных систем, для которых они изначально разработаны. То есть задача их интеграции и использования в рамках единой интегрированной ИС зачастую является невыполнимой. Это обстоятельство является важной отличительной особенностью таких подсистем, которая не позволяет осуществить полную интеграцию всех расчетных подсистем в рамках построения единой информационной системы, основывающейся на интеграции информации из информационных источников различных информационных систем по свойствам неорганических веществ.

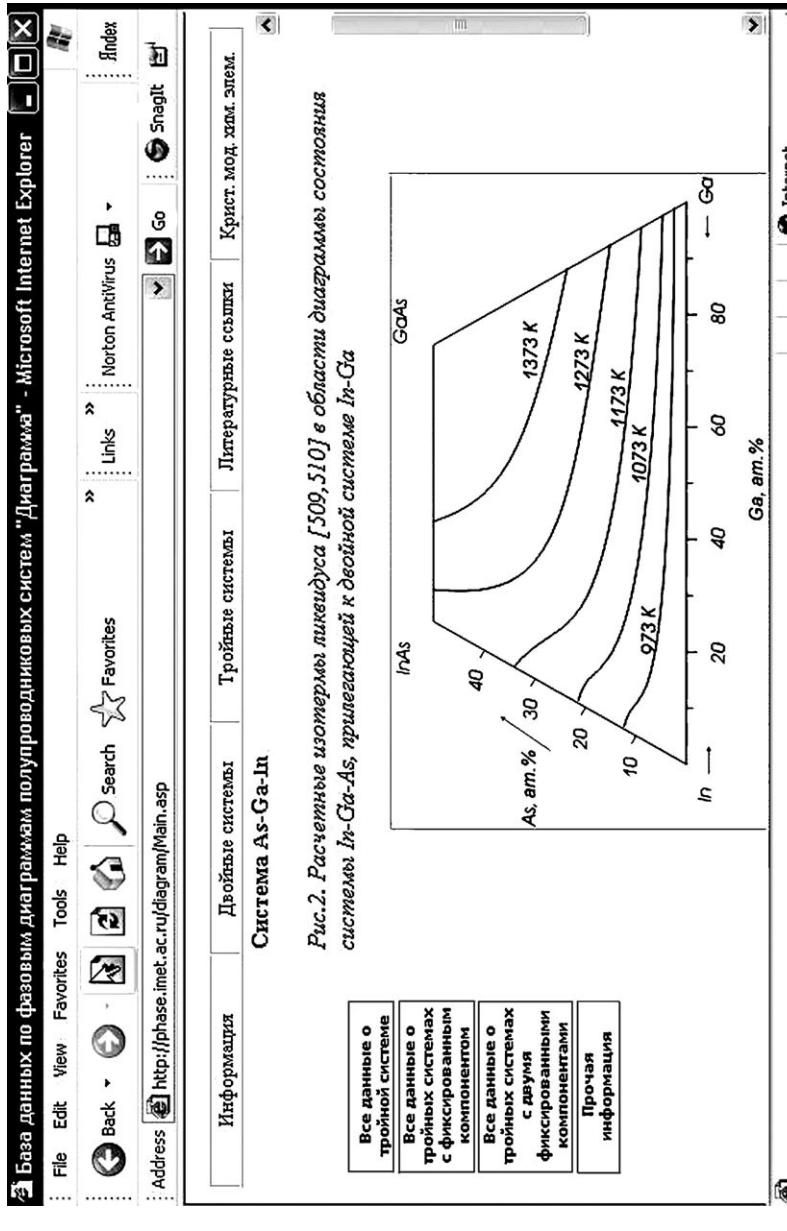


Рис. 2.5.1. Пример выдачи графической информации из БД «Диаграмма»

Краткие выводы

В главе получены следующие результаты:

- Выполнен обзор ИС СНВМ для электронной промышленности.
- Формализована иерархия понятий, используемых при построении информационных моделей данных при интеграции ИС СНВМ.
- Разработана ИС по информационным ресурсам неорганической химии «IRIC» (http://iric.imet-db.ru).
- Рассмотрены архитектуры современных ИС СНВМ и выявлены системные проблемы, затрудняющие применение созданных комплексов при компьютерном конструировании неорганических соединений.
- Предложена архитектура интегрируемых ИС СНВМ для информационной поддержки при компьютерном конструировании неорганических соединений.
- Разработана ИС «Кристалл» по свойствам акустооптических, электрооптических и нелинейнооптических веществ (русско- и англоязычные версии: http://crystal.imet-db.ru).
- Разработана ИС «Bandgap» по ширине запрещенной зоны используемых в электронике неорганических веществ (http://bg.imet-db.ru).

Глава 3
Системный подход к интеграции информационных систем

3.1. Методы интеграции гетерогенных информационных систем

3.1.1. Актуальность интеграции

В настоящей работе интегрируются информационные системы (ИС) по свойствам веществ и технологий их получения, и в первую очередь — ИС по свойствам неорганических веществ, используемых в электронике.

Прогресс электроники, как наиболее динамически развивающейся отрасли высоких технологий, в значительной степени обусловлен использованием новых веществ. В связи с этим, актуальным является решение проблемы обмена информацией между разработчиками и потребителями веществ, используемых в электронной технике. Традиционная система публикации результатов научных разработок — статья, затем обобщение в виде монографии или справочника — не соответствует высоким темпам развития электроники, элементная база которой обновляется каждые полтора-два года. Существенным фактором, усложняющим поиск необходимой для специалистов информации, является разбросанность данных по многочисленным литературным источникам разного профиля.

Современная информационная система для научных работников и инженеров, использующих вещества для электроники, должна обеспечивать оперативность обновления данных, их достоверность и полноту, а также возможность доступа к информации из глобальной сети Интернет. Именно эти принципы положены в основу разработанной в настоящей работе распределенной системы баз данных по свойствам веществ для электронной техники.

Следует отметить, что во всем мире огромные средства тратятся на нужды интеграции информационных систем. Так, уже в 2002 году затраты на интеграцию информационных систем и оценку качества данных по всему миру составили порядка одного триллиона долларов США [37]. Следует также отметить, что по данным Forrester 33 % всех компаний

в сфере ИТ занимаются интеграцией информационных систем [38]. А обзор за февраль 2002 года, проведенный CIO Magazine, показывает, что самой высокоприоритетной статьей расходов многие ИТ-компании считают построение интегрированных систем (рис. 3.1.1).

Рис. 3.1.1. Наибольшие расходы ИТ-компаний по сферам деятельности

3.1.2. Проблемы при интеграции информационных систем

Создание централизованной информационной системы, как правило, является сложной задачей даже в рамках одной крупной научно-исследовательской организации. Это обусловлено использованием различных информационных комплексов для сбора и регистрации данных, а также спецификой и разнообразием исследований. Поэтому проблема создания систем интеграции информации, которые бы были способны объединить всю важную информацию, накопленную исследователями данной организации, является актуальной при создании практически любой централизованной информационной системы.

Основной задачей при разработке централизованных систем является задача стандартизации. Стандартизации подвергаются все подсистемы, входящие в состав централизованной системы. В свою очередь, стандартизация подсистем и информационных потоков между ними осуществляется на основе собранной информации о взаимодействии всех составных частей, образующих информационную систему.

Следует отметить, что информация в различных информационных системах может храниться не только в форме распространенных баз дан-

ных, но и в других видах. Примером могут служить электронные таблицы (например, Microsoft Excel), CSV (Comma-Separated Values или другие ASCII flat-file), данные в формате XML, бинарные структуры данных, специально разработанные для хранения информации [40]. Все это значительно затрудняет интеграцию информационных систем.

В случае баз данных, использующих различные СУБД, возникает масса трудностей, а именно:

- Базы данных, использующиеся в различных организациях, построены на основе различных СУБД (Microsoft SQL Server, Oracle, IBM DB2 и т. д.).
- Базы данных всегда имеют различную структуру (схему БД) и оперируют различными данными.
- Репликация баз данных, требующая полного переноса данных из одной БД в другую, зачастую затруднительна по техническим и организационным причинам.

3.1.3. Методы интеграции ИС

Задача интеграции информации в настоящее время является актуальной для многих организаций, поскольку позволяет повысить эффективность их работы. Этим объясняется большой интерес к данному направлению развития ИТ, и появление множества новых программных продуктов от крупнейших компаний, направленных на решение задач интеграции. Проблема же, однако, заключается в том, разные компании по-разному понимают интеграцию и, следовательно, по-разному подходят к решению задач интеграции. Следует отметить, что это происходит на фоне еще не вполне четко сформировавшегося, размытого терминологического аппарата. Таким образом, необходимость разъяснения сути методов интеграции и их преимуществ привела в июле 2001 года к созданию лидерами в области интеграции международного консорциума по интеграции (Integration Consortium — IC). Следует отметить, что до мая 2004 года у консорциума IC было другое название — консорциум отрасли интеграции корпоративных приложений (EAI Industry Consortium — EAIIC), которое было изменено, поскольку консорциум занимался всеми вопросами интеграции, а EAI является лишь одним из методов интеграции. В настоящее время IC — это международная некоммерческая организация, объединяющая в своих рядах более 50 компаний из различных стран мира. В работе IC принимают участие не только поставщики программного и аппаратного обеспечения и системные интеграторы, но и потребители методов интеграции, представители научных кругов. Поскольку IC задумывался как сообщество, целью которого является единение отрасли интеграции, все

члены консорциума могут совместно определять проблемы и разрабатывать решения. По сути роль консорциума IC в сфере интеграции эквивалентна роли консорциума W3C в области Web-технологий.

В данной работе будем стараться придерживаться термина «метод интеграции» вместо «технологии интеграции», т. к. его использование является более уместным при разработке методологии интеграции. Методология рассматривается как система методов исследований, в данной работе — методов (или технологий) интеграции. А метод является набором методик, т. е. совокупностью приемов практической реализации.

В настоящее время выделяют три метода интеграции. Это интеграция корпоративных приложений (Enterprise Application Integration, EAI), интеграция корпоративной информации (Enterprise Information Integration, EII) и программное обеспечение для извлечения, преобразования и загрузки данных (Extract, Transform, Load — ETL) [39].

Принципы интеграции, заложенные в этих методах, используются для решения широкого круга задач: от интеграции в режиме реального времени до пакетной интеграции, и от интеграции данных до интеграции приложений. На рис. 3.1.2 показано положение названных методов по отношению к этим двум типам задач. Для интеграции данных в режиме реального времени лучше всего подходит подход EII. Для пакетной интеграции данных — ETL. А для интеграции приложений, в режиме реального времени или пакетном, наиболее подходящим инструментом является применением метода EAI. Следует отметить, что ни один из существующих на сегодняшний день методов интеграции не способен решить все проблемы, возникающие при объединении ИС [39].

Как было отмечено, в настоящее время происходит не только становление терминологической базы в области интеграции, но и развитие самих интеграционных подходов. Вследствие этого наблюдается некоторая неоднозначность в отношении того, каковы функции каждого из трех описанных

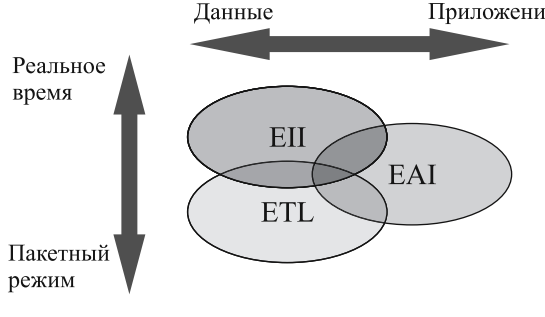


Рис. 3.1.2. Современные методы интеграции ИС [39]

методов, и в каких случаях тот или иной метод должен использоваться. Необходимо четко представлять возможности каждого метода и определить класс задач, для решения которых подходит каждый из методов. Для понимания различий в назначении методов интеграции необходимо привести соответствующие определения, учитывающие их назначения [41]:

- **EAI** — это метод интеграции, с помощью которого организация добивается централизации и оптимизации интеграции корпоративных приложений, обычно используя те или иные формы оперативной доставки информации (push technology), которая управляется внешними событиями (event-driven);
- **ETL** — это метод интеграции, который преобразует данные (обычно с помощью их пакетной обработки) из операционной среды, включающей гетерогенные технологии, в интегрированные, согласующиеся между собой данные, пригодные для использования в процессе поддержки принятия решений. Метод ETL ориентирован на консолидацию разнородных базы данных в виде, например, хранилища данных, витрины или операционного склада данных;
- **EII** — это метод интеграции в режиме реального времени несопоставимых типов данных из многочисленных источников как внутри, так и за пределами организации. Инструменты EII обеспечивают универсальный уровень доступа к данным и используют технологию поиска информации (pull technology) или возможности работы по запросам.

Для более полного понимания этих методов необходимо рассмотреть их взаимосвязь в рамках уже существующей информационной инфраструктуры организации. Для простоты приводится схема использования трех указанных подходов в организации, имеющей две информационные системы, которые необходимо интегрировать.

На рис. 3.1.3 показано, как каждый из этих методов может быть использован наилучшим образом. Метод EAI интегрирует транзакции двух или более приложений, метод ETL интегрирует данные операционных систем и компонентов поддержки принятия решений, а метод EII осуществляет виртуальную интеграцию данных из различных источников.

Рассмотрим более подробно сценарии использования каждого из указанных методов интеграции. Метод EAI используется, когда необходимо связать информационные системы в реальном времени. Раньше для обозначения подобного рода интеграции часто использовался термин Business Process Integration (интеграция бизнес-процессов — BPI). Если же речь идет об интеграции ИС разных организаций, то такую интеграцию часто называют B2B-интеграцией (Business-to-Business). Метод EAI применяется также в ситуации, когда необходимо, чтобы изменения, внесенные в одну ИС (обычно это небольшой набор записей), были отражены во всех

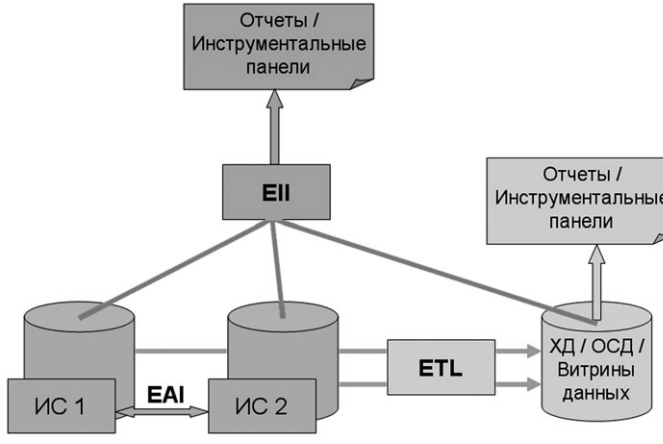


Рис. 3.1.3. Способы интеграции двух информационных систем с использованием методов интеграции EII, ETL, EAI [41]

других ИС. Этот метод используется при решении задач фиксации изменений и их переноса в соответствующие ИС и часто применяется, например, в банковской сфере.

Интеграционный метод ETL оказывается наиболее полезным в том случае, когда необходимо создать хранилище данных (ХД), содержащее хорошо документированные и надежные данные для исторического анализа, например, для анализа временных рядов или многомерных запросов. Этот метод также используется для интеграции ключевых справочных данных. ETL-метод незаменим для таких задач, как удаление дублирующихся данных, осуществление процессов проверки качества данных и т. п. ETL-инструменты также используются для создания отдельных витрин данных, обслуживающих конкретное подразделение организации или предназначенных для каких-либо долгосрочных целей. Инструменты ETL предоставляют пользователю возможность запустить повторяющиеся процессы для большей слаженности действий и возможности их многократного использования. Такие процессы включают создание точных технических метаданных, поддерживающих общую целостность среды Business Intelligence (BI).

Метод интеграции EII предназначен для случаев, когда необходимо создать общий шлюз (gateway) с едиными языком и точкой доступа к несогласованным источникам данных. Такие инструменты предоставляют приложениям и конечным пользователям возможности более гибкого, а также незапланированного доступа к данным. Стоит отметить, что при

этом не требуется постоянного использования данных или долговременных целей для получения этого доступа. Помимо традиционных реляционных баз данных, инструменты EII могут работать с XML и LDAP-источниками, плоскими файлами (ASCII flat-file) и другими нереляционными источниками информации. Инструменты EII являются особенно полезными, если есть необходимость добавить к справочным данным хранилища данных дополнительные детали, например, детальную информацию в режиме реального времени (например, сопоставление исторических данных с текущей ситуацией).

Следует отметить, что кроме понимания того, когда необходимо использовать эти методы, нужно также знать и проблемы, которые им присущи. Во-первых, внедрение этих методов интеграции требует от ИТ-персонала глубокого понимания тех требований, которые предъявляются к данным для принятия как тактических, так и стратегических решений. Применительно к методу ETL это означает, что необходимые данные извлекаются, преобразуются и загружаются в виде, пригодном для использования непосредственно аналитиками, системами поддержки принятия решений (СППР) или EII-сервером. В случае EII-интеграции способы представления данных должны удовлетворять требованиям аналитиков, предъявляемым к построению отчетов, т. е. данные должны быть пригодны для использования в аналитических отчетах. Во всех случаях понимание структур источников данных и требований, предъявляемых к данным, является необходимым шагом при внедрении этих методов интеграции и, безусловно, оправдывает то время, которое приходится тратить, чтобы достичь этого понимания.

Кроме того, необходимо понимать, что внедрение этих инструментов в уже сложившуюся архитектуру требует от ИТ-персонала разработки такой стратегии управления данными и ИС, которая будет постоянно поддерживать этот процесс в активном состоянии. Обязательной составляющей такой стратегии должно быть осознание того, что повышается важность механизмов архивирования, а также того, что с самого начала должны быть созданы контрольные журналы. Это необходимо для обеспечения слаженности и надежности интегрированных данных и приложений.

Следует отметить, что очень важен мониторинг производительности и эффективности описанных методов интеграции в условиях конкретной инфраструктуры. Их производительность в значительной степени будет зависеть от скорости архивирования данных, объемов и детальности данных, а также от эффективности функционирования ИС в условиях полной нагрузки. При определении производительности также следует оценить влияние, которые эти инструменты могут оказывать на операционные приложения и системы. Поэтому необходим постоянный мониторинг и этого влияния.

Учитывая то, что одной их основных целей, которые ставились при разработке интегрированной ИС по свойствам неорганических веществ, являлось использование интегрированной ИС для систем поддержки принятия решений (СППР), необходима интеграция информационных источников по свойствам неорганических веществ. Из приведенного выше краткого обзора методов интеграции очевидно, что в качестве таких интеграционных подходов на уровне данных целесообразно использовать EII и ETL, которые позволяют извлекать требуемые данные из информационных источников при создании интегрированной ИС.

EII (Enterprise Information Integration — интеграция информации на уровне предприятия или интеграция корпоративной информации) как термин был предложен в мае 2002 года Aberdeen Group. Информационные системы, построенные с применением принципов EII, обеспечивают универсальный доступ к множеству источников данных без предварительной их загрузки в хранилище данных. Такие системы были названы системами интеграции данных. С момента появления этого метода, как в технологии, так и на рынке были созданы несколько EII-продуктов и накоплен значительный опыт применения [42].

В настоящее время можно выделить несколько факторов, стимулирующих развитие EII-индустрии:

- Технологии управления и обработки информации, разработанные исследователями, находятся уже в достаточно зрелом состоянии и могут реализовываться для дальнейшего использования в информационных системах.
- Изменилось отношение и потребности организаций к управлению своими данными. Например, появилась необходимость создавать Web-сайты, требующие интеграции данных из множества различных источников. Помимо этого, тесная интеграция исследований и экономики вынуждает организации взаимодействовать друг с другом различными способами, интегрируясь в общие процессы, что невозможно без обеспечения единого информационного пространства.
- Появление XML как универсального механизма представления данных подталкивает к обмену информацией.
- Решения, основанные на методе хранилищ данных, кажутся неподходящими для решения поставленных задач, так как стоимость разработки специализированных систем постоянно возрастает и становится непозволительной (в частности, разработка программных посредников и адаптеров данных).

Следует отметить, что архитектура, использующаяся во всех программных продуктах, основана на схожих технологических принципах.

Сценарий интеграции информации условно состоит их нескольких этапов [42]:

- выявление источников информации, которые будут интегрированы;
- разработка виртуальной схемы (также часто называемой медиаторной схемой), которая будет использована конечными пользователями для построения запросов к интегрированной информации;
- программная реализация интегрированной ИС с учетом принятых ранее решений.

Обработка запросов начинается с так называемого переформулирования запросов к виртуальной схеме в запросы к исходным источникам данных. Затем следует процесс выполнения запросов с помощью специального обработчика, который создает план для выполнения отдельных подзапросов и последующего объединения их результатов с учетом возможностей и ограничений каждого источника.

Рассмотрим этот процесс более подробно. Метод EII использует запросы для сбора и интеграции данных и контента из многочисленных источников. EII-запрос является интегрированным, так как он сформулирован на основании интегрированного отображения источников данных. Для того чтобы выполнить такой запрос, EII-сервер опрашивает источники данных, находит релевантные данные и обрабатывает их в контексте приложения. Ранее создатели EII использовали упрощенный подход — все релевантные данные извлекались из источников данных в XQuery-процессор и полностью там обрабатывались. Для того чтобы понять, почему данный подход не позволяет достичь оптимальной производительности, рассмотрим пример запроса, требующего выполнения операции соединения (join query) двух очень больших таблиц из двух разных источников данных. Каждая таблица должна быть сначала преобразована в XML-формат (при этом ее размер возрастает примерно втрое), перемещена по сети, а затем соединена с другой таблицей. При этом возникает сразу две проблемы с точки зрения производительности. Первая заключается в том, что огромный объем информации будет постоянно передаваться по сети. Вторая заключается в том, что операция соединения пока еще остается не оптимизированной в разработанных XQuery-процессорах. Вместо этого следует применять технологии параллельной обработки и оптимизации запросов.

Так как источники данных располагаются на компьютерах, имеющих различные аппаратные платформы, операционные системы и СУБД, необходимо произвести декомпозицию поступающего к EII единого запроса на составные части, которые и будут пересылаться к источникам данных. Затем результаты этих подзапросов должны быть «собраны» с помощью специализированного ПО. В некоторых EII-продуктах роль этого специа-

лизированного ПО выполняет специальная РСУБД (например, IBM Information Integrator) или XQuery-обработчик. Компонентные запросы обычно выполняются через адаптеры данных (data wrappers) и направляются непосредственно к этому источнику. Адаптеры данных представляют собой специализированное ПО, учитывающее особенности конкретного источника данных. Характер компонентных запросов зависит от свойств схемы интеграции данных, которая участвует в переводе интегрированной схемы в схемы источников данных, и от оптимизатора запросов EII.

Критические аспекты производительности EII-систем связаны с распределенной архитектурой программного обеспечения EII. А именно, производительность будет зависеть от следующих параметров: (1) максимизации параллельности обработки запросов и (2) минимизации объемов данных, необходимых для сборки ответа на запрос и выбора наилучшего по производительности места для сборки этого ответа. Эти проблемы уже были достаточно хорошо решены разработчиками параллельных серверов баз данных, и схожие методы могут быть использованы и архитекторами EII-систем.

Следует отметить, что оптимизация запросов для EII — сложная проблема. К сожалению, вследствие конкуренции производители вынуждены выпускать на рынок ПО EII в крайне сжатые сроки, что часто приводит к упрощенным решениям, которые не способны масштабироваться и давать высокую производительность.

В настоящее время некоторые компании начинают построение систем с использованием модели данных XML и языка запросов XML. Эти компании вынуждены решать проблемы интеграции в двойном объеме, так как исследования по эффективной обработке запросов и интеграции для XML еще только начинаются. Следует отметить, что в настоящее время XQuery находится только в стадии становления, т. к. W3C выпустила окончательную версию спецификации XQuery 1.0 лишь в январе 2007 года [43]. Консорциумом W3C стандартизованы XPath 2.0 и XSLT 2.0, активно используемые при разработке информационных систем, но их пока явно недостаточно, чтобы вывести системы EII на качественно новый уровень.

Самыми первыми системами, в которых успешно были применены методы интеграции данных, были CRM-системы (Customer Relationship Management). Основной задачей подобных систем являлось предоставление специалистам компании так называемого *глобального отображения* клиента, то есть всей сводной информации по конкретному клиенту, которая находилась во множестве различных источников. Такие системы были предназначены для обеспечения актуальной информацией о клиенте, и, как следствие, подобные системы вынуждены были строить запросы к различным источникам и отслеживать все их изменения в режиме реального времени.

Как и любая новая отрасль индустрии, EII столкнулась со многими задачами, некоторые из которых до сих пор серьезно препятствуют росту данной технологии. Вот наиболее значимые из данных задач:

- масштабирование и производительность;
- горизонтальный или вертикальный рост;
- интеграция со средствами EAI и другим ПО промежуточного уровня;
- управление метаданными и семантической гетерогенностью.

Рассмотрим кратко эти основные задачи.

Масштабирование и производительность. Обеспечение масштабируемости и производительности информационных систем всегда является взаимосвязанной задачей. Необходимо обеспечить приемлемое время отклика системы при росте количества интегрируемых источников информации. Проблема состоит в следующем — насколько эффективно обработчик запросов может разбить запрос пользователя на подзапросы в режиме реального времени, опрашивая при этом распределенные источники данных и обеспечивая адекватный ответ? В этом контексте инструментальные средства EII часто значительно уступают уже зрелому методу хранилищ данных. Метод EII изначально рассматривался как технология получения ответов на запросы в режиме реального времени. Тем не менее, с увеличением мощности современных компьютеров пока еще более медленные методы EII становятся более конкурентоспособными, а при грамотном построении EII-систем разница в скорости может быть практически нивелирована уже сейчас.

Горизонтальный или вертикальный рост. При разработке программного обеспечения EII необходимо изначально определить стратегию разработки. С одной стороны, можно построить горизонтальную платформу, которая достаточно универсальна и может быть использована в любом приложении. С другой стороны, можно разрабатывать специализированное ПО, предназначенное для удовлетворения потребностей определенной вертикали, т. е. максимально полно обеспечить решение задач одного класса. Аргументом для выбора вертикальной модели роста является то, что можно получить *полное* решение всей задачи, пусть и достаточно узкой. Аргументом для выбора модели горизонтального роста является то, что получаемая информационная система является наиболее общей, и, следовательно, может использоваться для решения более широкого класса задач. Следует отметить, что горизонтальный подход выбирают также и при невозможности точного выбора вертикальной разработки. В конечном счете, решение вопроса о выборе подхода сводится к тому, как оптимально разделить ресурсы (то есть речь идет о принятии решений при ограниченности ресурсов в условиях неопределенности).

Интеграция со средствами EAI и другим ПО промежуточного уровня. Следует отметить, что промежуточное программное обеспечение (middleware) является достаточно сложным, так как взаимодействует с несколькими уровнями другого программного обеспечения (ПО). К тому же разработчики подходят по-разному к решению задач, и зачастую трудно определить, какую именно часть задачи решает то или иное ПО. Появление множества EII-инструментов от различных компаний еще более усугубляет положение. Гораздо более зрелым сектором является EAI (Enterprise Application Integration — интеграция приложений на уровне предприятия), где ПО предназначено для подключения к другим приложениям с целью эффективного взаимодействия. Таким образом, EAI фокусируется на приложениях, а EII концентрируется на данных и запросах к ним. При этом при обработке запросов используются инструменты EII, а при обновлении данных — инструменты EAI. Отсюда следует, что разделение между инструментами EII и EAI может являться лишь временным явлением, и в будущем эти методы будут объединены. Прочие программные продукты, связанные с этими отраслями, обеспечивают инструменты для очистки данных (перед их помещением в хранилище данных), составления отчетов и анализа информации. Интеграция таких инструментов с EII и EAI стала бы значительным усовершенствованием указанных технологий.

Управление метаданными и семантической гетерогенностью. Одними из ключевых вопросов, возникающих при выполнении проектов по интеграции данных, является *определение местонахождения* (locating) и *понимание* (understanding) интегрируемых данных. Зачастую обнаруживается, что данные, необходимые для какого-либо приложения интегрированной системы, вовсе не содержатся ни в одном источнике данных организации. В других случаях требуются значительные усилия для того, чтобы понять семантическую взаимосвязь между источниками и выработать схему их подключения к централизованной системе. Инструментальные средства, позволяющие решать подобные задачи, находятся в зачаточном состоянии. Им необходимы как структуры для хранения и работы с метаданными на уровне предприятия, так и инструменты, которые смогли бы облегчить преодоление препятствий семантической гетерогенности между источниками информации и поддерживать эту инфраструктуру в режиме реального времени.

Следует отметить, что, несмотря на сложности становления, индустрия EII действительно существует. Так, по данным [42], в 2005 году доходы индустрии EII составили порядка 0,5 млрд долларов США. Вместе с тем очевидно, что используемый EII-инструментарий постоянно развивается, что позволяет на современом этапе реализовать полный потенциал EII-подхода к интеграции.

У находящейся фактически в зачаточном состоянии EII существует уже достаточно зрелая альтернатива в лице *метода хранилищ данных* (*ETL*). Еще в начале 90-х годов У. Инмон определил хранилища данных (Data Warehouse) [44] как «предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки управления».

За последние десятилетия метод хранилищ данных и связанные с ними ETL-технологии (Extract, Transform, Load — извлечение, трансформация, загрузка) добились признания и теперь широко используются для интеграции и анализа больших объемов разнородных данных. Эти технологии прошли успешный путь от загрузки ежемесячных дампов оперативных данных, слегка очищенных и немного трансформированных пакетными программами, до систем сложного уровня, управляемых метаданными, которые перемещают огромные массивы информации от мест их первоначального сбора до операционных хранилищ данных или витрин данных [45].

Потребность в снижении цен на готовые ИТ-решения и необходимость обработки запросов в режиме реального времени приводит к разработке и апробации методов интеграции информации согласно требованиям конкретных проектов. Например, анализ и составление отчетов в режиме реального времени является необходимым для получения конкурентных преимуществ на рынке при управлении предприятием. Именно эта задача, в конечном счете, привела к подходу «виртуальных хранилищ данных», которые позволили объединить разрозненные данные. Такого рода интегрированные системы и явились предвестниками EII. Сейчас на рынке появилось несколько EII-технологий, способных по требованию (on-demand) объединить разрозненные источники данных без их полного перемещения (репликации) и предоставляющих единый SQL- или XQuery-интерфейс к этим множественным источникам.

При появлении новой технологии часто считается, что она автоматически делает существующие методы устаревшими, и несмотря на то, что метод EII находится лишь в начале своего развития, он должна заменить хранилища данных. Следует отметить, что это мнение ошибочно, и привести два аргумента против него:

- Для того чтобы быть жизнеспособными, технологии EII должны быть настолько же производительными и масштабируемыми, какими, например, являются на сегодняшний день технологии РСУБД. Поэтому для того, чтобы адекватно и объективно измерить производительность, необходим стандартный тест на производительность, подобный тестам TPC для СУБД [46].
- Даже если EII-технологии будут полностью доработаны, они все же не заменят хранилищ данных. В зависимости от задач, которые нужно

решить, интегрируемые данные будут перемещаться либо в хранилище данных, либо к ним будет осуществлен виртуальный доступ с использованием EII-технологий [45].

Таким образом, предположение о том, что EII заменит хранилища данных, поверхностно и неверно. Хранилища данных будут продолжать использоваться для построения больших систем и решения сложных аналитических задач. Метод EII предназначен для быстрой и гибкой интеграции «по требованию». Следовательно, вопрос заключается в том, в каких случаях целесообразно хранить интегрируемую информацию в одном месте, а в каких — осуществлять доступ ко многим источникам для организации виртуального хранилища. Можно предложить несколько простых правил, которые можно применять при принятии решения.

Правила централизованного хранения данных в рамках хранилищ данных [45]:

- **Сохраняйте данные, чтобы хранить историю.** Хранилища данных хранят исторические (архивные) данные в том смысле, что они передаются из источников данных через определенные интервалы времени. Поскольку история больше нигде не хранится — хранилища данных должны использоваться для хранения архивных данных.
- **Сохраняйте данные централизованно, когда доступ к источникам информации затруднен или запрещен.** По многим причинам (организация работы, безопасность и т. п.) технологиям объединения БД может быть запрещен доступ к источникам данных. В этом случае данные из этих источников должны извлекаться в некоторое постоянное хранилище, такое как хранилище данных.

Правила виртуализации данных (объединения)

Эти правила должны использоваться только тогда, когда не срабатывает ни одно из правил централизованного сохранения данных [45]:

- **Виртуализировать необходимо данные, находящиеся за пределами границ хранилищ данных, а также новые витрины данных.** Вместо избыточного копирования данных по нескольким хранилищам данных необходимо виртуализовать совместное использование данных несколькими хранилищами. По определению, подобные измерения могут совместно использоваться несколькими витринами данных. Другим сценарием, в котором виртуальная витрина данных будет хорошим решением, является интеграция нового внешнего источника данных, который не был включен в начальную структуру данных хранилища с этим хранилищем или витриной данных.
- **Виртуализировать данные для особых проектов и для того, чтобы построить прототипы систем.** Данные можно довольно быстро собрать для одноразовых отчетов или для апробации прототипов новых

приложений посредством построения виртуальной схемы требуемых данных.

- **Виртуализировать данные, которые должны отражать *текущие* факты.** В приложениях, таких как инструментальные панели или порталы, данные должны отражать сиюминутное состояние операций. Для приложений подобного класса интеграция с использованием EII является необходимой и единственно возможной, в отличие от хранилищ данных, отражающих данные с различной степенью задержки.

Следует подчеркнуть, что главным отличием между EII и методом хранилищ данных (ETL) является то, что EII — метод интеграции данных «по требованию» (или технология «вытягивания» данных из источников в режиме реального времени), в то время как ETL — метод, при котором разрозненные данные помещаются в единое информационное хранилище данных заранее и, фактически, полностью готовы к запросам [47]. Таким образом, одним из главных преимуществ EII над ETL является то, что пользователь, вследствие гибкости технологии EII, получает доступ к *текущему* состоянию данных (так называемые live data — «живые данные»). Существует множество сценариев, когда это действительно необходимо, например, резервирование товаров на складах и оформление заказа.

Если учесть, что новые источники данных, удовлетворяющие общей схеме, могут быть динамически обнаружены интегрированной системой и подключены к ней, то у EII появляется еще одно преимущество по сравнению с ETL — *способность динамической интеграции* с источниками данных. Следует отметить, что в ETL данные из нескольких заданных (и тем самым фиксированных) информационных источников по определенным правилам помещаются в централизованное хранилище данных.

Со стороны стоимости решения, преимуществом EII является то, что нет необходимости в хранилище или другом виде репозитория для скопированных данных, как в ETL. Но с другой стороны, объем хранилища, его стоимость и время обработки информации, необходимое, чтобы наполнить его, предсказуемы. В то время как при использовании EII возникает масса вопросов, например, непредсказуемость времени загрузки и общей производительности интегрированной информационной системы. Однако, необходимо учитывать, что в этом случае ETL-подход не является панацеей, так как подчас просто невозможно загрузить *все* релевантные данные в одно хранилище данных.

Необходимо подчеркнуть, что потенциально EII предлагает большую гибкость и более содержательный подход к интеграции данных, чем метод хранилищ данных. Однако для того, чтобы работать достаточно быстро, аппаратные платформы, на которых функционирует EII, должны быть основаны на высокопроизводительных серверах БД с параллельной обработкой

потоков информации, а не пытаться заменить их. Трансляция схемы и декомпозиция объединенных запросов должны быть нацелены на генерирование компонентных запросов, которые могут быть переадресованы достаточно высокопроизводительным серверам баз данных для эффективного исполнения. Следует отметить, что проблема выбора метода интеграции — это проблема постановки задачи. Каждый из рассмотренных выше методов (EII, ETL, EAI) эффективен для решения различных классов задач, поэтому только в редких случаях эти методы могут заменять друг друга.

3.1.4. Проблемы при интеграции гетерогенных источников информации

Ответ на запросы при создании интегрированных систем влечет за собой множество различных конфликтов, которые, в общем, могут быть названы конфликтами гетерогенности. Конфликты можно условно разделить на следующие классы [61]:

- **Платформенные и системные** — интегрируемые системы используют несовместимые аппаратные платформы, операционные системы, СУБД и другое программное обеспечение для их функционирования.
- **Синтаксические и структурные** — интегрируемые системы используют разные по синтаксическому описанию (XML, RDF, реляционные таблицы) и по структуре (реляционные данные, объекты) данные, т. е. отличия в моделях данных и их схемах.
- **Семантические** — разные источники данных для обозначения одной и той же сущности могут использовать различные значения. Например, для обозначения типа кристаллической структуры перовскита в различных источниках могут использоваться: «перовскит», «$CaTiO_3$», «perovskite» и т. д. В дополнении к конфликтам обозначений (naming conflicts), могут встречаться и конфликты шкал и точности (scaling & precision conflicts). Так, например, значения температуры могут быть указаны в градусах по Цельсию и Фаренгейту в разных источниках данных с разной точностью.

Для того чтобы согласовать гетерогенные представления интегрируемых данных, необходимо задать соответствующие правила отображения (mapping rules). При этом следует отметить, что знание предметной области интегрированной системы является необходимым для успешного решения конфликтов гетерогенности и оптимизации работы системы (оптимизация запросов).

При этом нужно учитывать, что при построении интегрированной системы всегда необходимо искать компромисс между требованиями, предъяв-

ляемыми к интегрированной системе, ибо никакая система не может одинаково хорошо обеспечивать следующее:

- поддержку часто обновляемых источников данных;
- частое изменение потребностей пользователей в предоставляемой информации;
- использование неопубликованных данных в исходных источниках данных.

Таким образом, разработка интегрированной системы — это всегда компромисс между ее простотой, функциональностью и стоимостью. При разработке интегрированной ИС по свойствам неорганических веществ требуется успешно разрешить все три класса конфликтов гетерогенности.

3.2. Системный анализ методов интеграции

3.2.1. Базовые информационные процессы в локальных ИС

Типовая структура ИС включает в себя ряд подсистем, реализующих базовые информационные процессы сбора, хранения, передачи, обработки и представления информации. На рис. 3.2.1 представлены информационные процессы, протекающие в локальной ИС. В ней реализуются все основные процессы (кроме информационного обмена с внешними ИС).

Запрос от пользователя, сформированный при помощи интерфейса (1) поступает в модуль управления, который на основе метаданных (2) обращается к подсистеме хранения данных. Далее выполняется непосредственное извлечение (3) и обработка данных (4). Результаты отображаются пользователю при помощи интерфейса (5).

Переход от локальной БД к распределенной, но однородной БД требует минимальных изменений в схеме обработки информации. Метабаза должна быть дополнена сведениями о распределении данных по множественным источникам. Наличие гетерогенных ИС, обладающих различными форматами хранения данных и различными процессами их обработки, обуславливает необходимость модификации процессов обмена информацией и требует применения того или иного метода интеграции ИС.

Создание централизованной информационной системы, как правило, является сложной задачей даже в рамках одной крупной научно-исследовательской организации. Это обусловлено использованием различных информационных комплексов для сбора и регистрации данных, а также спецификой и разнообразием исследований. Поэтому проблема создания систем

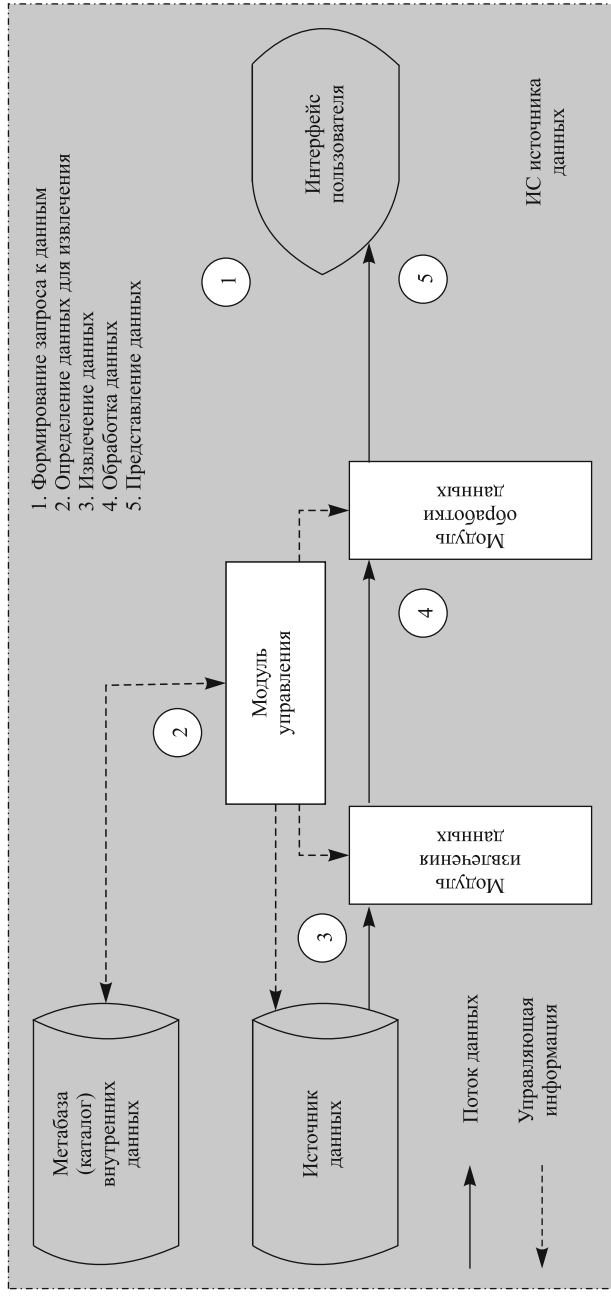


Рис. 3.2.1. Информационные процессы в локальной ИС

интеграции информации, которые бы были способны объединить всю важную информацию, накопленную исследователями данной организации, является актуальной при создании практически любой централизованной информационной системы.

При переходе к интегрированным ИС необходимо в первую очередь ответить на следующие вопросы:

- Какие подсистемы интегрированной ИС будут распределенными, а какие останутся локальными?
- Какие подсистемы интегрированной ИС станут (изначально или в перспективе) гетерогенными, а какие останутся однородными?
- Каков будет баланс между централизацией и периферийностью в системе управления интегрированной ИС?

В первом приближении можно сказать, что методы EII и ETL основаны на использовании источников данных, а метод EAI предполагает распределенную обработку сообщений.

Выбор метода интеграции определяет характеристики, которыми будет обладать интегрированная система. В контексте конкретной задачи по интеграции, характеристики могут являться как недостатками, так и положительными свойствами, позволяющими решить данную задачу наиболее оптимально и эффективно.

Основной задачей при разработке централизованных систем является задача стандартизации. Стандартизации подвергаются все подсистемы, входящие в состав централизованной системы. В свою очередь, стандартизация подсистем и информационных потоков между ними осуществляется на основе собранной информации о взаимодействии всех составных частей, образующих информационную систему [62].

3.2.2. Метод интеграции корпоративной информации EII

Интеграция корпоративной информации — это интеграция данных из многочисленных систем в унифицированное, согласованное и точное представление, которое предназначено для изучения и обработки данных.

При организации процесса интеграции данных по технологии EII главным функциональным модулем является «предметный посредник» (иногда называемый модулем извлечения), который обеспечивает:

- Единый интерфейс взаимодействия конечных приложений с источниками исходной информации.
- Поиск запрашиваемой информации по исходным базам данных.
- Агрегацию собранной информации для передачи конечным приложениям.

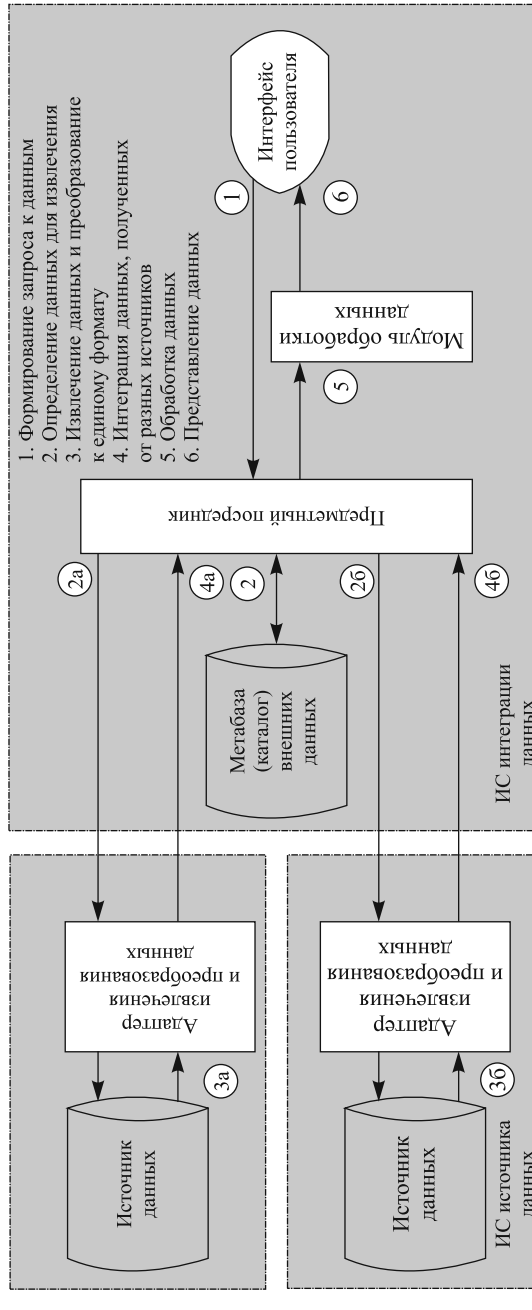


Рис. 3.2.2. Метод интеграции корпоративной информации ЕП

Взаимодействие с источниками хранения исходных данных осуществляется за счет адаптеров — модулей преобразования форматов данных.

Схема интеграции разнородных источников данных на основе метода интеграции корпоративной информации представлена на рис. 3.2.2.

Конечные приложения инициируют запросы, определяющие характер и объем интегрируемых данных. Для взаимодействия между предметным посредником и приложениями используется единый, стандартизированный в рамках данной системы интеграции данных, интерфейс для прикладных программ (Application Programming Interface, API).

Предметный посредник определяет, к каким источникам данных необходимо обратиться для получения запрашиваемой информации. Источники данных определяются на основе информации, содержащейся в метабазе — специальном каталоге, содержащем описание информации, находящейся в источниках исходных данных.

Определив источники информации, предметный посредник оправляет контекстные запросы индивидуально к каждому источнику исходных данных. Формат запросов стандартизирован и одинаков для всех источников данных. Для конвертации запроса в формат взаимодействия с источником данных используется индивидуальный адаптер.

После извлечения (pull), данные агрегируются и передаются конечным приложениям. На этапе агрегации возможно преобразование и изменение данных, устранение конфликтов данных.

С точки зрения конечного приложения взаимодействие осуществляется с единой базой данных в едином стандартизированном формате.

3.2.3. Метод интеграции на основе хранилищ данных ETL

Название метода ETL является аббревиатурой от названий функций извлечения (Extract), преобразования (Transform) и загрузки (Load) данных.

Интеграция разнородных источников данных включает в себя предварительное формирование хранилища данных и последующую работу с данными, размещенными не в ИС источников данных, а в хранилище данных.

Формирование хранилища данных состоит из трех этапов.

На первом этапе интегрируемые данные извлекаются из источников данных (source), в качестве которых могут выступать любые организованные хранилища данных. Метод извлечения зависит от структуры и технической реализации источника. Может быть использовано прямое подключение (native connection) к базе данных, запросы к системе (message querying), программный интерфейс (API) и т. д.

Взаимодействие является однонаправленным — при извлечении данных инициатором выступает система синхронизации. Извлечение производится в пакетном режиме — через заданные временные интервалы, которые могут зависеть от множества факторов, включая частоту обновления данных источника и человеческий фактор, и отличаться для каждого отдельного источника.

При первичном извлечении данные извлекаются из базы данных источника в полном объеме. При последующих извлечениях данных для оптимизации работы системы может быть реализован механизм определения изменений данных источника и извлечения только данных, необходимых для актуализации информации в промежуточном хранилище (Staging Area).

В результате выполнения первого этапа интеграции по методу ETL система интеграции локально сохраняет данные, полученные от источника, в промежуточном хранилище и может применить функции преобразования данных.

На втором этапе, с помощью функций преобразования выполняется унификация представления данных промежуточных хранилищ для создания единой структуры хранения и организации данных. На данном этапе выполняются функции объединения и слияния или, наоборот, разделения данных; изменения формата представления данных — например, реорганизация таблиц и отношений между таблицами; добавление новых атрибутов; сортировка и фильтрация. Также осуществляется анализ и контроль качества и полноты собранных данных, устраняются конфликты интеграции данных.

По завершении данного этапа, информация в промежуточных хранилищах приводится в единый формат, определяющий взаимодействие сформированной базы данных с инструментальными панелями и программным обеспечением.

На третьем этапе осуществляется загрузка данных в постоянное хранилище (интегрированных) данных. Хранилище данных (warehouse) содержит непосредственно данные и метабазу данных.

После выполнения функций загрузки данных формуется база интегрированных данных, имеющая единую детерминированную структуру и интерфейс, с помощью которого любые модули и приложения могут обращать информации, хранимой в базе.

Функции ETL могут одновременно применяться к нескольким базам данных источников, либо к группам баз данных источников, в случае однотипности последних.

Схема интеграции разнородных источников данных на основе хранилища данных представлена на рис. 3.2.3.

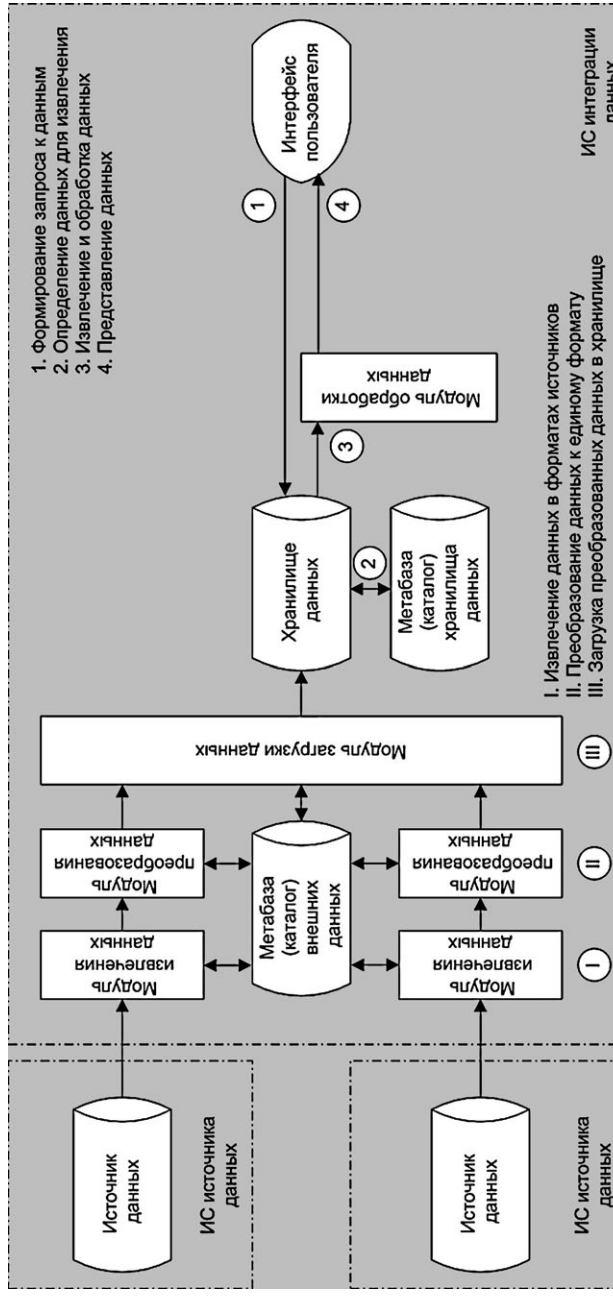


Рис. 3.2.3. Интеграция на основе хранилищ данных ETL

Последующая работа с хранилищем данных не отличается от работы с локальной базой данных. В ИС интеграции на основе хранилища данных реализуются все базовые информационные процессы (рис. 3.2.1) обработки информации.

3.2.4. Интеграция корпоративных приложений EAI

Метод интеграции корпоративных приложений EAI вместо непосредственной интеграции разнородных данных предполагает интеграцию результатов работы двух и более приложений (программ), работающих с независимыми друг от друга данными.

Метод EAI позволяет автоматизировать процессы работы с разнородными данными без необходимости непосредственного обращения к данным и изменения готовых интерфейсов, программ и приложений работы с данными.

Основной задачей в контексте данного метода интеграции является задача организации взаимодействия между объединенным интерфейсом работы с приложениями и приложениями-источниками — согласования формата, средств и способов передачи данных от одного приложения к другому.

Существует несколько наиболее распространенных методов решения данной задачи:

- использование программных адаптеров (Adapters) для обоих приложений;
- использование промежуточного программного обеспечения, ориентированного на обработку сообщений (Message-oriented middleware, MOM);
- использование репликатора данных (Data Replication Engine).

Программный адаптер является модификацией приложения, обеспечивающей прием/передачу данных в формате понятном как приложению-источнику, так и объединенному интерфейсу. Реализация адаптера зависит от конкретного приложения.

Использование промежуточного ПО обеспечивает синхронизацию информации между приложениями с помощью запросов, передаваемых в асинхронном режиме. Формат передаваемых между приложениями сообщений также должен быть согласован.

Использование репликаторов обеспечивает синхронизацию данных на уровне баз данных. При этом непосредственная интеграция приложений не осуществляется. Репликатор отслеживает изменения в базе-источнике и в случае обнаружения изменений — передает их базе данных, взаимодействующей с объединенным интерфейсом.

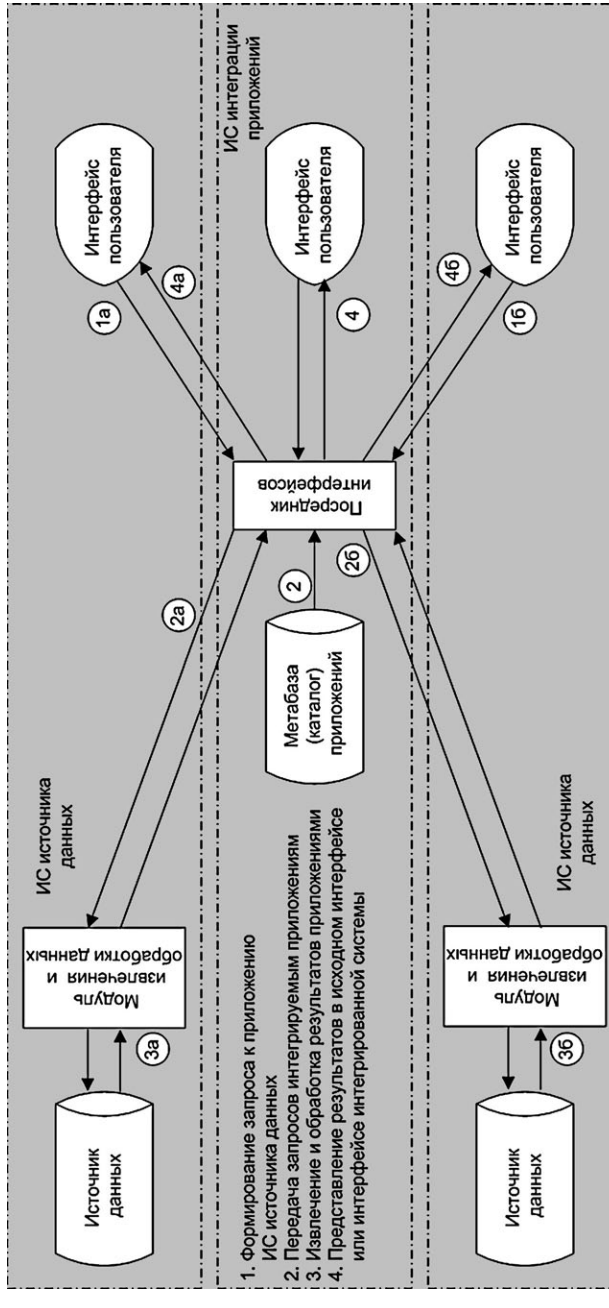


Рис. 3.2.4. Интеграция корпоративных приложений EAI

Схема интеграции разнородных источников данных на основе метода интеграции корпоративных приложений представлена на рис. 3.2.4.

Иногда при использовании меотда интеграции корпоративных приложений EAI дополнительно уточняется, какие именно корпоративные приложения имеются в виду — относящиеся к одной корпорации или к разным. В рамках одной организации интеграция корпоративных приложений обычно описывается термином Business Process Integration (BPI — интеграция бизнес-процессов). Если же речь идет об интеграции ИС разных организаций, то такую интеграцию часто называют B2B-интеграцией (Business-to-Business).

3.2.5. Обобщенная схема методов интеграции гетерогенных информационных систем

Появление каждого из описанных выше методов интеграции обусловлено необходимостью решения определенного круга задач, которые, независимо от отрасли или характера деятельности возникали перед компаниями и организация с ростом объемов используемых данных и расширением ИС.

В ряде случаев возможно использование единственного варианта интеграции данных. Например, отсутствие доступа к исходным данным предопределяет использование метода интеграции приложений EAI, а требование доступности данных независимо от работоспособности ИС источника данных — применение метода хранилищ данных ETL.

В табл. 3.1 приведены критерии сравнения методов интеграции гетерогенных ИС для подбора наиболее подходящего варианта реализации интеграции для каждого конкретного случая.

При объединении ИС информационные процессы 1–5 (рис. 3.2.1) будут реализованы в различных ИС (множественных ИС источников данных либо в центральной ИС интеграции) при помощи специализированных программных компонентов (модулей). На основе системного анализа информационных потоков составлена обобщенная схема интеграции гетерогенных ИС (рис. 3.2.5). Пунктиром на схеме показаны условные границы интегрируемых ИС.

ИС источников данных могут работать автономно в локальном режиме (верхняя часть схемы). Интеграция приложений EAI требует применения в посредника интерфейсов, управляющего передачей сообщений между интегрируемыми приложениями на основе метабазы внешних приложений. При этом извлечение и обработка данных выполняются в ИС источников данных, а результаты могут быть представлены как в интерфейсе ИС интеграции, так и в интерфейсах исходных ИС.

Интеграция на основе метода хранилищ данных ETL включает модули извлечения исходных данных в форматах ИС источников (на основе метабазы внешних данных), преобразование их к формату хранилища данных и загрузки в локальное хранилище (на основе метабазы хранилища данных). Локальное расположение всех модулей обработки данных требует доступности ИС источников только на момент первичного извлечения данных.

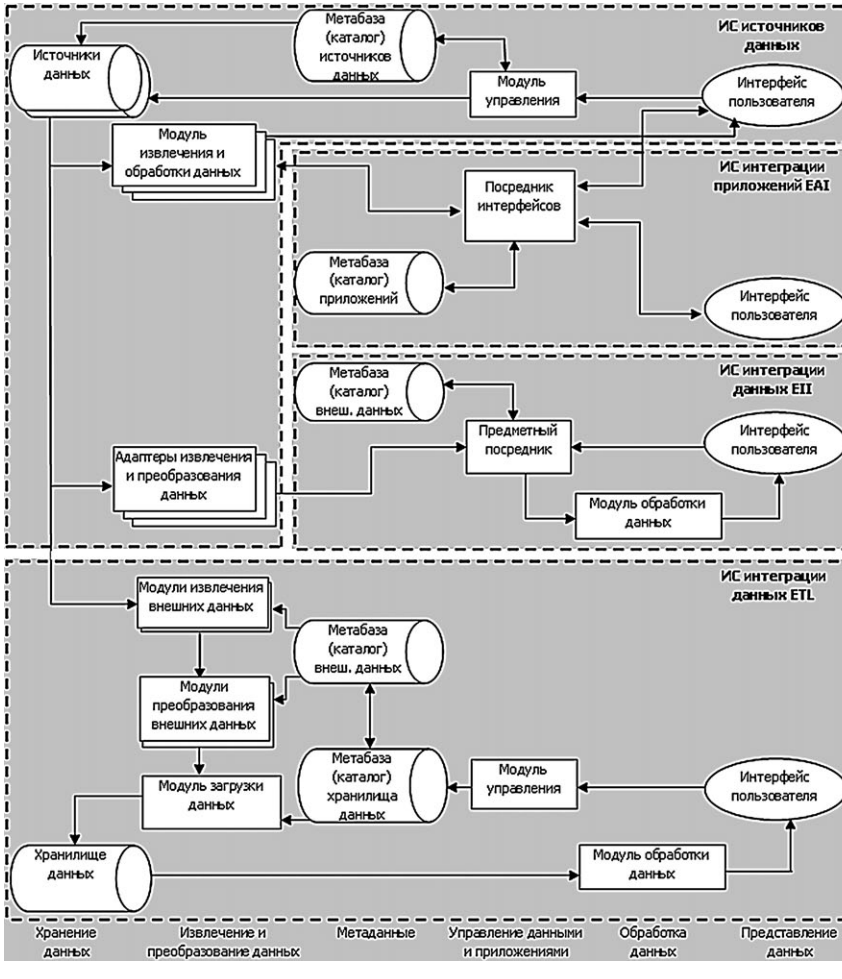


Рис. 3.2.5. Обобщенная схема интеграции гетерогенных ИС

Таблица 3.1. Критерии сравнения методов интеграции гетерогенных информационных систем

Критерий	Локальные БД	ETL	EM	EAI
Объект интеграции		Исходные данные	Исходные данные	Приложения, обрабатывающие исходные данные
Объем извлекаемых данных	Только запрашиваемые пользователем данные	Все данные	Только запрашиваемые пользователем данные	Только запрашиваемые пользователем данные
Доступ к данным источника	Требуется, частичный в момент запроса к данным	Требуется, в полном объеме в момент извлечения данных	Требуется, частичный в момент запроса к данным	Не имеется
Актуальность извлекаемых данных	Данные всегда актуальны	Актуальны на момент последней загрузки	Данные всегда актуальны	Данные всегда актуальны
Хранение извлеченных данных	Долговременное, в собственном хранилище данных	Долговременное, в собственном хранилище данных	Кратковременное, в оперативной памяти	Кратковременное, в оперативной памяти
Формат извлекаемых данных	Определяется ИС источника данных	Определяется ИС источника данных	Определяется ИС интеграции данных	Определяется ИС источника данных
Извлечение данных	Выполняет ИС источника данных	Выполняет ИС интеграции данных	Выполняет ИС источника данных	Выполняет ИС источника данных
Преобразование формата данных		Выполняет ИС интеграции данных	Выполняет ИС источника данных	Выполняет ИС источника данных
Обработка данных	Выполняет ИС источника данных	Выполняет ИС интеграции данных	Выполняет ИС интеграции данных	Выполняет ИС источника данных
Представление данных	Выполняет ИС источника данных	Выполняет ИС интеграции данных	Выполняет ИС интеграции данных	Выполняет ИС источника данных и/или интеграции приложений

При использовании метода интеграции данных EII исключается трудоемкая стадия разработки и заполнения промежуточного хранилища данных, но требуется постоянный доступ к ИС источников данных и размещение в исходных ИС адаптеров извлечения данных и преобразования к единому формату ИС интеграции.

При интеграции гетерогенных ИС (в отличие от локальной ИС) необходима реализация процессов внешнего информационного обмена. На обобщенной схеме интеграции (рис. 3.2.5) эти процессы представлены стрелками информационных потоков, пересекающими условные границы ИС. Также процессы передачи информации имеют место при реализации удаленного доступа пользователей к интерфейсу ИС интеграции.

В результате анализа критериев сравнения методов интеграции (табл. 3.1) и обобщенной схемы интеграции гетерогенных ИС (рис. 3.2.5) можно определить ряд ситуаций, в которых использование одного конкретного метода интеграции является предпочтительным, либо единственно возможным. Рекомендации по выбору предпочтительного метода интеграции гетерогенных ИС приведены в табл. 3.2.

Так, если непосредственный доступ к данным ИС источника отсутствует, то использование методов интеграции данных EII и ETL невозможно, а единственным доступным способом является интеграция приложений.

Таблица 3.2. Рекомендации по выбору предпочтительного метода интеграции гетерогенных ИС

Критерий принятия решения по выбору метода интеграции	**Условия интеграции гетерогенных информационных систем**	**Рекомендуемый метод интеграции**
Возможность доступа к данным источника	Доступ к данным отсутствует	EAI
	Доступ к данным возможен	ETL или EII
Надежность доступа к данным источника	Необходим постоянный доступ	ETL
	Постоянный доступ не требуется	EII
Хранение извлеченных данных	Необходимо локальное хранение	ETL
	Не требуется	EII или EAI
Интеграция расчетных подсистем ИС	Требуется	EAI
	Не требуется	ETL или EII
Ограниченность доступа к данным источника	Доступ на ограниченной (платной) основе	EII или EAI
	Возможен полный доступ	ETL
Актуальность извлекаемых данных	Требуется	EII или EAI
	Не требуется	ETL

Постоянный доступ к данным может быть обеспечен (не считая локальных БД) только при использовании метода хранилищ данных ETL. Работоспособность интегрированной ИС на основе методов EII и EAI зависит от доступности ИС источников данных.

Требование локального хранения данных может быть вызвано не только необходимостью обеспечения постоянного доступа к ним, но и целым рядом других причин, например, для организации собственной системы разграничения доступа к данным (по соображениям безопасности, на платной основе и т. д.).

Наличие патентованных (или недоступных по другим причинам) алгоритмов обработки данных ограничивает выбор только методом интеграции приложений EAI, поскольку создание равноценного приложения обработки извлеченных данных (в рамках интегрированной ИС) по вышеуказанным причинам является невозможным.

Невозможность полного доступа к данным ИС источника исключает применение метода хранилищ данных. Платный доступ к данным ИС источника определяет высокую стоимость хранилища данных и делает его разработку экономически неэффективной.

Метод хранилищ данных предполагает локальное хранение не только полного объема исходных данных, но и различных промежуточных данных (в процессе их преобразования для загрузки), поэтому ограниченность ресурсов хранения исключает применение этого метода.

Метод хранилищ данных ETL предполагает также определенную периодичность выполнения процедур извлечения внешних данных и загрузки преобразованных данных в локальное хранилище ИС интеграции. Если эти процедуры являются трудоемкими, дорогими, осложнены частой сменой внешних форматов данных и т. д., то это часть может приводить к возможной потере актуальности загруженных в хранилище данных.

Обеспечение полной актуальности данных может быть достигнуто только за счет использования методов EII, либо EAI. Кроме того, преобразование данных в этих методах осуществляется в рамках ИС источников данных. Таким образом, смена форматов исходных данных отражается на интегрированной ИС в минимальной степени.

Использование метода хранилищ данных (ETL) предлагается для создания интегрированного источника данных в рамках одной организации, например, ИМЕТ РАН. Это позволит получить максимальную надежность и скорость работы с интегрированными данными со стороны систем компьютерного конструирования неорганических соединений или других высокоуровневых средств интеграции. Использование метода интеграции данных (EII) предлагается для виртуальной интеграции материаловедческой информации между ИС, как правило, относящимися к разным организациям, запрещающим физическое копирование данных или предоставляю-

щими ограниченный доступ к данным на платной основе. Таким образом, на нижнем уровне (в рамках организации) данные интегрируется с помощью хранилищ данных (ETL), а затем на более высоком уровне интеграция осуществляется с использованием метода EII (рис. 3.2.6). Отмечается, что возможна реализация многоуровневой схемы использования хранилищ данных и виртуальной интеграции для обеспечения требуемой скорости обработки и масштабируемости.

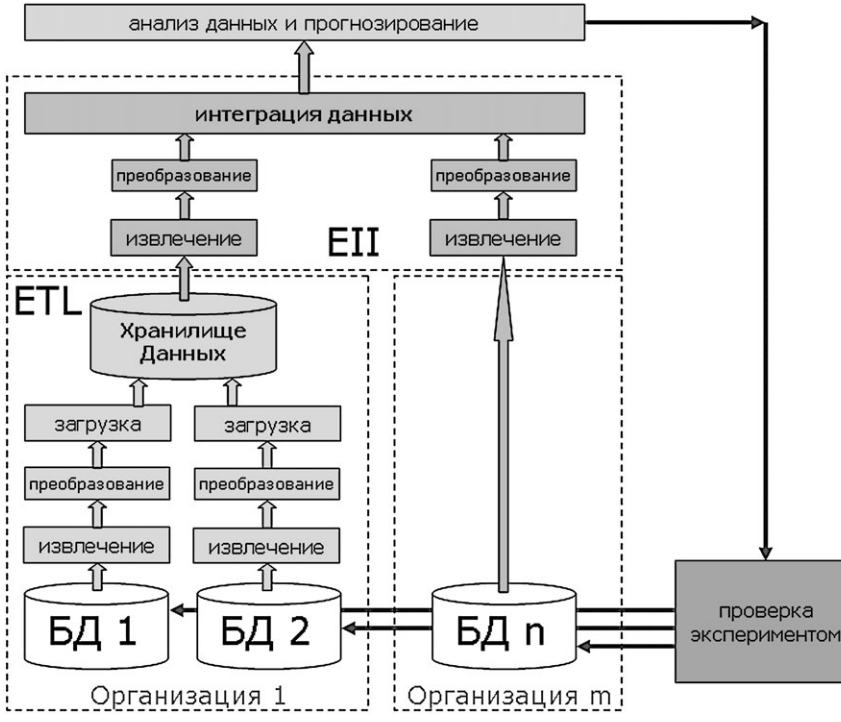


Рис. 3.2.6. Методика консолидации данных ИС СНВМ

3.3. Методология интеграции информационных систем

Отличительной особенностью ИС СНВМ, интегрируемых в настоящей работе, является то, что все они, как правило, являются предметно-ориентированными и поэтому хранят информацию только о тех веществах

и их характеристиках, которые относятся к исследуемой предметной области. Например, ИС по фазовым диаграммам систем с полупроводниковыми фазами «Диаграмма» и ИС по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами «Кристалл» — проблемно ориентированы на специалистов в области химии и электронной техники.

Таким образом, в разных информационных системах представлены различные характеристики (будем далее называть их свойствами) различных сущностей. Значения свойств определяются, в первую очередь, составом неорганических веществ (набором химических элементов, входящим в состав соединений, и их соотношением), а также в большинстве случаев физические свойства зависят от кристаллической структуры, т. к. в указанных выше ИС содержится информация о твердых фазах. Поскольку ИС тесно связаны с химией, то сущности в ИС описываются с помощью иерархии понятий (система → вещество → модификация) в виде дерева (см. рис. 3.3.1).

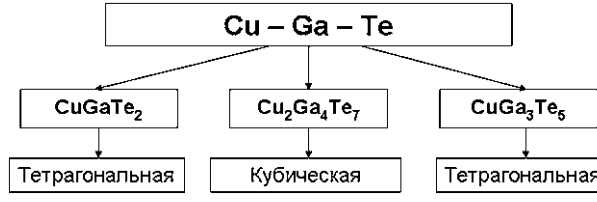


Рис. 3.3.1. Иерархия понятий

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 объектов можно использовать три уровня: система, вещество и кристаллическая (полиморфная) модификация (далее — модификация). Приведем кратко определения этих терминов [155]:

Химическая система (элементы, определяющие качественный состав) — система, образованная химическими элементами. Она может быть описана как множество атомов, образующих химическую систему. Более строго, химическая система — 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 Cu–Ga–Te.

Химическое соединение — однородное вещество постоянного или переменного состава с качественно отличным химическим или кристалло-

химическим строением, образованное из атомов одного или нескольких химических элементов. Характерной особенностью химического соединения яв-ляется его однородность.

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 определения, как и все созданные человеком понятия, являются нечеткими (размытыми). В связи с этим иногда трудно провести границу между, например, упорядоченным твердым раствором и соединением т. п.

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 описываются на уровне химических веществ (например, температура плавления, растворимость и пр.), а некоторые свойства описываются на уровне конкретных модификаций (например, нелинейнооптические коэффициенты, коэффициенты Селмейера 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.

Очевидно, что, учитывая наличие в ИС по свойствам веществ разнородных гетерогенных данных и расчетных подсистем, использование которых возможно только в рамках исходных ИС, необходим комплексный подход к интеграции информационных систем. При этом интегрированная ИС, построенная по данному механизму, должна обеспечить конечного пользователя всей информацией, содержащейся в рамках объединяемых ИС. Для выработки такого подхода необходимо учитывать архитектуру современных ИС по свойствам веществ, что позволит наиболее эффективно решить задачу интеграции информационных систем [156, 284].

Как было показано выше при исследовании архитектуры современных ИС по свойствам веществ, серверная часть ИС разделена на две составляющие:

- база данных информационной системы (БД ИС);
- Web-приложение информационной системы (Web-приложение ИС).

В БД ИС содержится структурированная информация по определенному тематикой ИС разделу предметной области. Учитывая определенное родство предметных областей, рассматриваемых нами ИС по свойствам веществ, в них можно выделить общий структурный типаж, согласно которому определенным образом представляется вся информация, хранящаяся в различных гетерогенных БД. Поскольку в разных ИС рассматриваются различные свойства объектов, определенных спецификой предметной области, достаточно классифицировать эти объекты таким образом, чтобы информация из всех (или подавляющего большинства) БД в данной предметной области «укладывалась» в заданную модель представления информации.

Модель представления химических объектов может быть получена при анализе понятий предметной области с учетом структурной особенности различных БД по свойствам веществ. Как показывает анализ указанной предметной области, информация по свойствам объектов в различных ИС предметной области может храниться на следующих трех уровнях:

- уровень химических систем (определен качественный состав вещества);
- уровень химических веществ (определен количественный состав вещества);
- уровень кристаллических модификаций химических веществ.

Эта иерархия понятий описывает самый верхний уровень структурной организации предметной области. В ней не учитываются способы получения и обработки соединений, методы их исследования и прочие важные параметры. Тем не менее, все описываемые в БД объекты предметной области могут быть сведены к одному из трех типов объектов предлагаемой нами классификации (системы, вещества или модификации). Необходимо также определить правила работы с этими объектами. Важнейшим моментом здесь является определение операции сравнения объектов из разных источников информации. Определение этой операции позволит интегрированной системе отличать разные объекты и находить тождественно равные объекты в различных интегрируемых ИС, что позволит аккумулировать описания свойств данного объекта из различных ИС. Таким образом, появится возможность работы с различными свойствами данного объекта, содержащимися в различных БД ИС. Возможным также станет агрегирование данных из различных источников информации, объединенных такой общей моделью.

Как уже было отмечено, в БД ИС содержится структурированная информация по разделу предметной области, освещаемому ИС. Причем стоит отметить, что пользователь обращается к этой информации не напрямую, а через посредника, которым в нашем случае является Web-приложение ИС. Именно оно отвечает за предоставление нужной пользователю информации, преобразуя данные из БД в понятную и удобную для пользователя форму.

Зачастую, наряду со структурированными данными, ИС содержит информацию в неструктурированном виде. Например, это могут быть аналитические обзоры, содержащие текстовое описание в произвольной форме той информации, которая не может быть структурирована в рамках существующей ИС. Широко распространены графические пояснения в виде рисунков и графиков, которые используются практически повсеместно. Часто в ИС включаются расчетные подсистемы, которые осуществляют вычисление каких-либо параметров и вывод результатов пользователю. Например, программа D_Mapper в ИС «Диаграмма» осуществляет подготовку информации о фазовых диаграммах для визуализации и масштабирования, информационно-расчетная подсистема для компьютерного моделирования процессов жидкофазной эпитаксии [113], разработанная в МИТХТ им. М. В. Ломоносова, осуществляет расчет и визуализацию процессов жидкофазной эпитаксии на основании вводимой пользователем информации и данных, содержащихся в БД ИС. Стоит отметить, что функционирование всех расчетных подсистем осуществляется в рамках Web-приложения ИС, то есть Web-приложение ИС является своего рода естественным интерфейсом к расчетным подпрограммам.

Итак, подводя некоторые итоги, следует отметить, что существует два класса методов интеграции современных ИС СНВМ [157]. Первый класс методов заключается в интеграции информационных ресурсов на уровне их гетерогенных источников информации (БД). При этом решается вопрос получения данных (в соответствии с заданными схемами) из разрозненных источников в рамках общепринятой модели данных, т. е. осуществляется реализация инфраструктуры для работы с интегрированными данными [158]. Эти данные затем могут быть выведены в необходимом формате пользователям интегрированной информационной системы или преобразованы к нужному формату с помощью дополнительных преобразований. Стоит отметить, что преобразованная к нужному формату и агрегированная из нескольких источников информация может являться входными данными для различных систем поддержки принятия решений (СППР), таких как программы компьютерного конструирования неорганических соединений.

Очевидно также, что для того, чтобы предоставить конечным пользователям доступ к богатым возможностям расчетных подсистем, входящих в состав соответствующих интегрируемых ИС по свойствам веществ, необходимо проводить интеграцию на уровне Web-приложений интегрируемых ИС [159, 160]. Таким образом, второй класс методов интеграции заключается в необходимости объединить не сами информационные источники (БД ИС), а только их пользовательские интерфейсы, из которых осуществляется доступ к информационно-расчетным подсистемам. Как уже было отмечено, такими интерфейсами являются Web-приложения соответ-

ствующих информационных систем. Этот способ, с одной стороны, позволяет не изменять коренным образом инфраструктуру каждой из отдельных информационных систем, а значит, и наладившуюся технологию администрирования данных (корректировки и добавления информации), а с другой стороны, позволяет конечному пользователю получить доступ сразу ко всему спектру информации о веществе, хранящейся в различных ИС СНВМ.

Таким образом, при нынешних условиях развития современных ИС СНВМ, необходима разработка методологии интеграции ИС СНВМ. Это означает необходимость использования интеграции ИС СНВМ как на уровне данных, содержащейся в гетерогенных информационных источниках ИС (ETL и EII), так и на уровне пользовательских интерфейсов ИС (метод EAI). При этом следует особо отметить, что обеспечивается независимость развития отдельных ИС СНВМ, так как созданные ИС СНВМ продолжают свое развитие, пополняясь информацией и расширяясь новыми расчетными подсистемами, обеспечивающими дополнительные функциональные возможности ИС СНВМ. Предложенная методология позволяет успешно сочетать в себе все достоинства разработанных и продолжающих эволюционировать ИС СНВМ (рис. 3.3.2). В рамках разработанной методологии предоставляется как доступ к текущим пользовательским

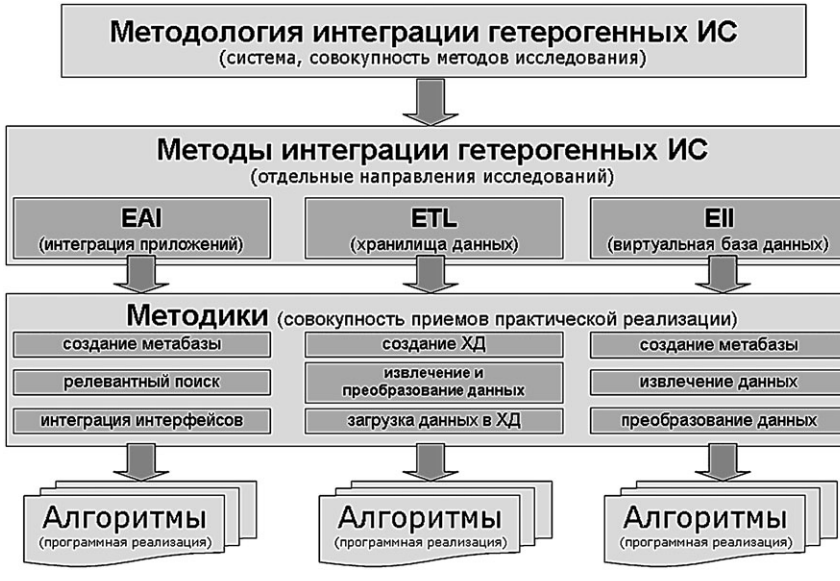


Рис. 3.3.2. Методология интеграции ИС СНВМ

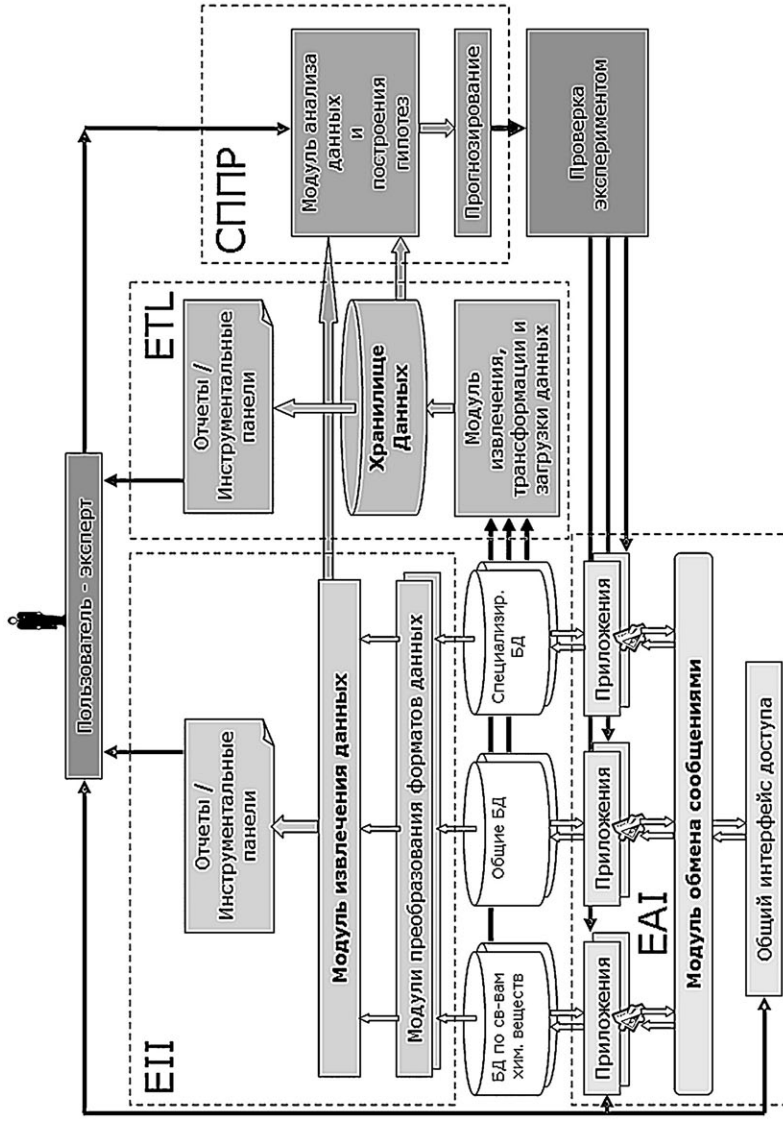


Рис. 3.3.3. Архитектура интегрированной ИС СНВМ и связь с СППР

интерфейсам ИС СНВМ и свободное перемещение пользователей между ними, так и богатые возможности по сбору и агрегации информации, полученной из гетерогенных распределенных источников данных по свойствам веществ, согласно общей разработанной информационной схеме [140].

На основе предложенной методологии и анализа информационных потоков при разных методах интеграции ИС СНВМ, учитывая связь интегрированной ИС с системами поддержки принятия решений (СППР) при исследовании и производстве химических соединений для электронной техники, предлагается архитектура построения интегрированной ИС СНВМ (рис. 3.3.3). Отличительная особенность предлагаемой архитектуры заключается в использовании известных методов интеграции для обеспечения максимально тесной консолидации ИС СНВМ. Э что позволит использовать разработанную ИС как конечным пользователям, так и программным комплексам компьютерного конструирования неорганических соединений.

3.4. Интеграция гетерогенных источников данных информационных систем

Проведение интеграции информационных систем по свойствам неорганических веществ на уровне источников данных является актуальным, поскольку дает возможность строить запросы ко всем виртуально объединенным источникам данных и извлекать из них информацию в режиме реального времени. Следует отметить, что в отличие от объединения Web-приложений (см. главу 6), где применяется EAI-интеграция, при объединении гетерогенных источников данных речь идет уже о EII- или ETL-интеграции. Тогда как объединение Web-приложений интегрируемых ИС актуально, прежде всего, для конечного пользователя, собирающего необходимую ему информацию. Объединение источников данных предназначено, главным образом, для использования в системах поддержки принятия решений (СППР), поскольку позволяет получить доступ к широкому спектру данных из объединяемых ИС.

Основой при построении интегрированных ИС в рамках методов интеграции данных является разработка общей глобальной схемы предметной области. Следует отметить, что, в отличие от подхода Global-as-View EII или ETL, при подходе Local-as-View EII глобальная схема должна быть особенно тщательно продуманной, поскольку именно в ее терминах будут представлены все интегрируемые источники информации ИС СНВМ. Таким образом, в эту схему должны быть достаточно полно заложены понятия предметной области, чтобы схема не являлась узким местом при описании

информационных источников и не искажала бы семантику информации, которая в них содержится. При этом глобальная схема должна быть не слишком громоздкой, чтобы чрезмерно не усложнять процесс манипуляции данными в контексте интегрируемой ИС. Как видно, задача построения глобальной схемы предметной области во многом является задачей достижения компромисса между точностью отображения понятий предметной области и простотой ее описания.

При проведении интеграции источников данных важно разработать схему интеграции, разрешающую три класса конфликтов гетерогенности: (1) платформенные и системные, (2) синтаксические и структурные и (3) семантические. Остановимся на способах разрешения этих конфликтов, которые предлагается использовать при построении интегрированной ИС СНВМ.

3.4.1. Разрешение платформенных и системных конфликтов

Суть платформенных и системных конфликтов заключается в том, что ИС, подлежащие интеграции, построены на различных аппаратных и программных платформах. Например, ИС работают под управлением разных операционных систем (Microsoft Windows, Unix и т. д.), СУБД (Microsoft SQL Server, Oracle и т. д.) и другого ПО (Web-сервера Microsoft IIS, Apache, nginx и т. д.). Таким образом, ИС строились на платформах, объединить которые еще совсем недавно представлялось крайне сложной задачей.

Огромные трудности, возникающие при попытках объединить ИС, разработанные на различных программно-аппаратных платформах, вынудили ИТ-сообщество искать пути интеграции различных архитектур, как программных, так и аппаратных. В результате осознания и формулирования проблем, возникающих при интеграции разрозненных ИС, начал формироваться стек стандартов, технологий и архитектурных шаблонов (design patterns), ориентированных на использование широко распространенной инфраструктуры Web и получивший название Web-сервисы [163].

Таким образом, благодаря стремительному развитию ИТ появилась технологическая возможность объединения ИС, построенных на различных программно-аппаратных платформах. Современный подход к обмену информацией между различными ИС основывается на использовании стандарта XML и Web-сервисов. Эти технологии стандартизированы консорциумом W3C и их реализации существуют для всех программно-аппаратных платформ. При этом важно отметить, что Web-сервисы не являются абсолютно новой парадигмой, призванной заменить существующие подходы и технологии интеграции. Web-сервисы стали естественным

развитием подходов интеграции, формализовавшим правила взаимодействия различных ИС с учетом требований работы в инфраструктуре глобальной сети Интернет.

Очевидно, что сам по себе стандарт Web-сервисов без соответствующей реализации со стороны современных технологических платформ бесполезен. В настоящее время, в области разработки и интеграции приложений доминируют следующие конкурирующие платформы — COM+/.Net, J2EE и CORBA. На сегодняшний день уже существует реализация технологии Web-сервисов на базе этих платформ. При этом следует отметить, что исторически COM+/.Net, J2EE и CORBA не обладали средствами взаимной интеграции, но функциональность Web-сервисов, представленная в новейших реализациях этих программных платформ, позволяет им не просто сосуществовать, но и прозрачно взаимодействовать друг с другом [164].

Особенно важно то, что Web-сервисы, являясь широко поддерживаемым ИТ-сообществом стандартом, продолжают свое развитие. Необходимо отметить формирование в 2002 году консорциума WS-I.org, основной задачей которого является контроль стандартов Web-сервисов для обеспечения совместимости различных реализаций. В рамках работ WS-I сформулирован так называемый First Profile, описывающий требования к реализациям Web-сервисов как набор стандартов, обязательных для реализации. К ним относятся SOAP, WSDL, UDDI. При этом необходимо понимать, что стандарты Web-сервисов сами по себе базируются на инфраструктурных стандартах работы в Интернете (например, HTTP и SMTP) и структурирования информации (XML, XML Schema [165]). Суть архитектуры приложений ИС, ориентированной на Web-сервисы, показана на рис. 3.4.1.

Web-сервисы основываются на XML и позволяют в настоящее время организациям интегрировать свои внутренние ИС, а также обеспечивать взаимодействие с внешними ИС. Основное преимущество технологии Web-сервисов по сравнению с другими технологиями интеграции заключается в том, что эта технология позволяет решить задачу легче, быстрее и с меньшими затратами. Следует отметить, что при интеграции ИС нет необходимости коренным образом пересматривать их архитектуру или переписывать исходный код. Необходимо лишь добавление к уже существующей ИС программного модуля, отвечающего за предоставление или потребление Web-сервисов. Таким образом, сохраняются уже сделанные инвестиции в существующие ИС СНВМ.

Технически Web-сервисы — это модульные приложения, имеющие стандартные интерфейсы для работы через Интернет. Другими словами, Web-сервисы являются слабосвязанными компонентами ПО, доступными для использования через стандартные протоколы сети Интернет, такие как HTTP и SMTP. Таким образом, Web-сервис — это программный сервис, к которому клиент может получить доступ и использовать его через сеть

Интернет. При этом Web-сервис функционирует по принципу черного ящика, то есть его внутренняя реализация полностью скрыта от клиента, его использующего. Для взаимодействия с Web-сервисом клиент должен только указать адрес Web-сервиса, который он хочет использовать. После этого он может вызывать методы требуемого Web-сервиса в соответствии с его спецификацией, описанной на языке WSDL, и тем самым получать требуемую функциональность. При этом клиенту не важно знать детали функционирования данного Web-сервиса. Он только вызывает Web-сервис и обменивается информацией, используя стандартный формат обмена сообщениями, оговоренный протоколом SOAP [163].

Не имеет смысла подробно останавливаться на технических деталях технологии Web-сервисов, отметим только, что она базируется на существующих открытых стандартах Интернет и стандартах, которые широко распространены или планируются к принятию в ближайшем будущем. Базовые стандарты, на которых основываются Web-сервисы — это HTTP (реже в качестве транспортного протокола используется SMTP), XML, SOAP, WSDL, UDDI и пр. Следует отметить, что поддержку Web-сервисов заявили фактически все ведущие мировые поставщики ПО: IBM, BEA, Sun, Microsoft, Oracle, Software AG, Borland и др. Именно это обстоятельство и позволяет использовать Web-сервисы, как способ межплатформенного взаимодействия, понятный всем современным технологическим платформам.

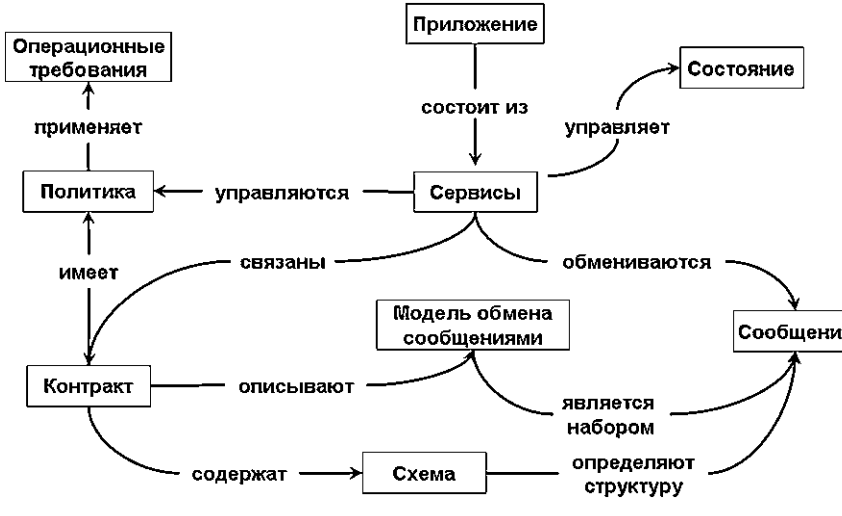


Рис. 3.4.1. Архитектура ИС, ориентированной на Web-сервисы

Таким образом, использование технологии Web-сервисов является оптимальным для построения интегрированной ИС по свойствам неорганических веществ, поскольку позволит объединить ИС, даже если они функционируют на различных программно-аппаратных платформах. Если вернуться к рассмотренной схеме интеграции источников информации ИС с применением подхода Local-as-View, то все оболочки источников данных (или адаптеры источников данных) должны быть выполнены в качестве Web-сервисов. При этом все эти Web-сервисы должны предоставлять доступ к конкретным информационным источникам согласно разработанной общей схеме. Иными словами, все Web-сервисы должны иметь одинаковое WSDL-описание, что обеспечит унифицированную работу со всеми такими Web-сервисными оболочками со стороны предметного посредника (медиатора) [185].

3.4.2. Разрешение синтаксических и структурных конфликтов

Как было отмечено выше, синтаксические и структурные конфликты возникают из-за того, что ИС используют различные по синтаксическому описанию и структуре данные. В ряде ИС используются реляционные СУБД, в других иерархические СУБД. В последнее время нередко строятся ИС, которые используют XML 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 и 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 по свойствам неорганических веществ взаимодействие с внешней программной средой не предусматривалось вовсе.

Разрешить синтаксические и структурные конфликты можно посредством введения общей схемы представления информации и обмена данными, построенной согласно описанию предметной области. Как уже было отмечено в разделе 3.3, при описании химических объектов можно использовать три уровня: система, вещество и кристаллическая модификация (далее — модификация). Указанная иерархия химических объектов, которая будет рассматриваться в контексте интегрированной ИС, представлена на рис. 3.4.2.

Таким образом, в общую схему предметной области закладывается три типа объектов, соответствующих химическим сущностям: элементы

(или химическая система — качественный состав вещества), вещество (количественный состав вещества) и модификация. Следовательно, все оболочки интегрируемых ИС должны оперировать этими тремя типами объектов при ссылке на химические сущности. При этом стоит учитывать, что если описывается определенная химическая модификация, то определена также и химическая система с веществом, модификация которого описывается. То есть если описание химической сущности ведется на уровне модификаций, то все вышележащие уровни (вещество и система) тоже описаны. Следует заметить, что обратное неверно. То есть, если описывается химическая система, то вещество и модификация четко не определены. Однако необходимо понимать, что при описании сущности на уровне системы все описанные свойства автоматически распространяются на все вещества и модификации, образованные этой системой. Это во многом напоминает наследование в объектно-ориентированном программировании (ООП).

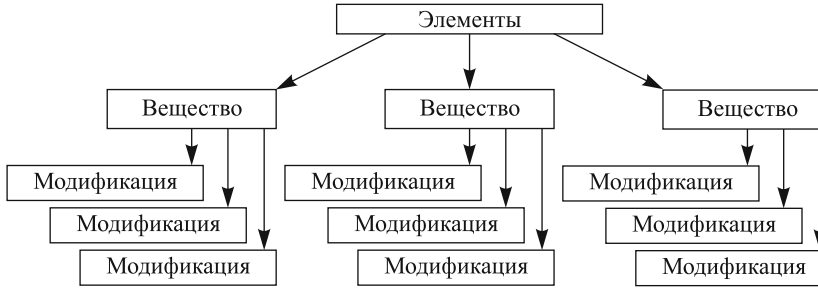


Рис. 3.4.2. Иерархия химических объектов, рассматриваемая в контексте интегрированной ИС СНВМ

Воспользуемся теорией множеств для описания сущностей нашей предметной области. Обозначим множество химических систем как S, множество химических веществ как C, а множество химических модификаций как M. Тогда химическая система будет обозначаться как s (где $s \in S$), химическое вещество обозначим через c (где $c \in C$), а химическую модификацию — как m (где $m \in M$).

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 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 химических элементов, которые образуют вещество, на множество пар $R^* \times R^*$, задающих соответственно минимальное и максимальное вхождение заданного химического элемента в вещество, раствор или смесь c. То есть $f : e_i \longrightarrow (R^*_{\min}, R^*_{\max})$, 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, $Fe_{1-x}Se$. $R^*_{\min}$ и $R^*_{\max}$, соответственно, минимальная и максимальная концентрация химического элемента e_i в веществе c. В случае, когда концентрация конкретного химического элемента e_i в веществе c фиксирована, то $R^*_{\min} = R^*_{\max}$. Химическая модификация m может быть представлена кортежем (s, f, mod), где $s \in S$, $f : e_i \longrightarrow (R^*_{\min}, R^*_{\max})$, а mod — строковое обозначение модификации вещества, принятое в интегрированной ИС.

Расширим множества C и M, добавив к ним пустой элемент $null$. То есть $null \in C, null \in M$. Учитывая то, что химическая сущность описывается только на одном из трех уровней (система, вещество и модификация) и при описании уровня все верхние уровни определены, а нижние неопределенны — null, любая химическая сущность (система, вещество и модификация) может быть описана тройкой (s, c, m), где $s \in S, c \in C, m \in M$.

Приведем примеры записи химических систем, веществ и их модификаций ($s \in S, c \in C, m \in M$):

- $(s, null, null)$ — шаблон для записи химических систем;
- $(s, c, null)$ — шаблон для записи химических веществ;
- (s, c, m) — шаблон для записи химических модификаций.

В шаблоны для записи химических объектов умышленно вносится избыточность, например в шаблоне тройки для записи химических веществ присутствует s, хотя точно такое же множество присутствует в кортеже c. Так же в шаблоне тройки для записи химических модификаций присутствует s и c, хотя эти сущности определены в кортеже m. Эта избыточность в запись вносится в целях наглядности, чтобы подчеркнуть наследование в иерархии химических объектов.

Как было отмечено, в интегрируемых ИС содержится информация по свойствам химических объектов. Например, плотность, растворимость, теплопроводность, ширина запрещенной зоны. При этом для каждой

химической сущности в БД ИС нередко содержится несколько записей для описания значения свойства. Это обусловлено несколькими обстоятельствами. Во-первых, информация, содержащаяся в БД ИС, может быть взята из различных источников, при этом данные нередко расходятся. Это объясняется как различными способами измерения, так и точностью измеряющей аппаратуры. Таким образом, в ИС приводится несколько вариантов значения, например, плотности различных кристаллов. Во-вторых, значения рассматриваемых свойств зачастую зависят от внешних условий, при которых приводились измерения. Например, такие параметры как растворимость и ширина запрещенной зоны зависят от температуры, при которой проводились измерения. 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 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, 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 даже отдельно взятого конкретного свойства, не говоря о представлении свойств в целом.

Таким образом, необходим некоторый механизм, позволяющий гибко представлять значения свойств в рамках интегрированной ИС. В настоящее время существует общепризнанное средство описания произвольных форматов данных — это XML (eXtensible Markup Language). С помощью этого языка разметки удобно описывать различные структуры данных, он является межплатформенным форматом и поддерживается большинством языков и библиотек [166]. На сегодняшний день именно этот язык является тем звеном, которое может служить основой для обеспечения взаимодействия различных программно-аппаратных платформ. Сейчас все большее количество информации в современных промышленных системах представляется в формате XML. Использование XML в качестве формата представления и обмена информацией является целесообразным еще и потому, что он используется как основа функционирования Web-сервисов.

Таким образом, для разрешения семантических и структурных конфликтов необходимо стандартизировать форматы представления описанных

химических сущностей и свойств в рамках интегрированной ИС на языке XML. То есть, необходимо разработать форматы соответствующих XML-документов для представления химических сущностей, их свойств и другой информации. Это позволит обмениваться информацией между звеньями интегрированной ИС.

3.4.3. Разрешение семантических конфликтов

Семантические конфликты во многом являются следствием синтаксических и структурных конфликтов, рассмотренных выше. Суть семантических конфликтов состоит в том, что разные ИС для обозначения одной и той же сущности используют разные обозначения. Это так называемые конфликты обозначений (naming conflicts). Следует отметить, что могут встречаться и конфликты шкал и точности (scaling & precision conflicts).

Нередко встречаются и конфликты точности, которые своим происхождением обязаны различиями в представлении чисел в различных ИС. Приведем пример с вариантами записи вещественного числа. Например, число 12345,6789 можно быть представлено в разных ИС так: 12345,6789 (тип данных numeric(18,10) в Microsoft SQL Server) или 12345,6787 (тип данных real в Microsoft SQL Server). Очевидно, что сами это числа не равны друг другу, хотя на самом деле имеется в виду одно и то же число. Поэтому если попытаться просто сравнить эти два числа между собой, то получится что они разные, хотя сущность имелась в виду одна и та же. Следовательно, в некоторых случаях имеет смысл закладывать некоторую погрешность и сравнивать числа с учетом этой погрешности. Т.е. числа a и b с учетом погрешности при сравнении ξ считаются равными тогда и только тогда, когда выполняется условие:

$$a = b \Leftrightarrow a \in [b - \xi * b, b + \xi * b],$$

где ξ — допустимая при сравнении погрешность.

Такой метод сравнения позволит «разглядеть» одинаковые числовые сущности. Главное, не сделать погрешность ξ при сравнении слишком большой, чтобы не получилось случая, когда пять равно трем (например, в большинстве случаев достаточно, чтобы $\xi = 10^{-6}$).

Для того чтобы согласовать гетерогенные представления интегрируемых данных, необходимо задать соответствующие правила отображения (mapping rules), которые позволили бы отыскивать идентичные информационные сущности в рамках разных ИС. При этом следует отметить, что знание предметной области интегрированной системы является необходимым для успешного решения семантических конфликтов гетерогенности.

В нашем случае необходимо обеспечить распознавание идентичных химических сущностей и их свойств в рамках интегрированной ИС.

Подавляющее большинство современных ИС используют внутренние идентификаторы (как правило, целочисленные) для идентификации сущностей внутри ИС. При этом естественным образом возникает ситуация, когда для обозначения одной и той же сущности разные ИС используют разные идентификаторы. Например, ИС «Кристалл» использует для обозначения вещества GaAs целочисленный идентификатор HeadClue = 82, а ИС «Bandgap» для обозначения того же вещества использует идентификатор SubstanceID = 137. Для обозначения модификации химических веществ в ИС используются разные строковые литералы. Например, в ИС «Кристалл» для обозначения триклинной модификации используется строковый литерал «тр», а в ИС «Bandgap» для обозначения той же модификации используется строковый литерал «Triclinic». Аналогичных примеров семантических конфликтов можно привести еще сколь угодно много.

Как уже отмечалось, для успешного разрешения семантических конфликтов необходимо знание предметной области и знание особенностей представления информации в интегрируемых ИС. Так как только специалист, обладающий не только знаниями предметной области, но и знанием особенностей построения ИС «Кристалл», сможет «расшифровать», например, строковый литерал «тр», обозначающий триклинную модификацию и связать его с другими строковыми литералами, обозначающими триклинную модификацию. Более того, в результате самого процесса сопоставления сущностей возможны ошибки, обусловленные недостаточно глубокими познаниями эксперта, отвечающего за процесс сопоставления. В нашем случае, например, эксперт может не знать, что строковый литерал «тр» соответствует триклинной модификации и не связать его с ней. Таким образом, может возникнуть ситуация, когда все химические модификации из ИС «Кристалл», помеченные строковым литералом «тр» будут являться самостоятельными сущностями, не связанными с триклинными модификациями других ИС. Необходимо построить интегрированную ИС так, чтобы эта разовая ошибка эксперта могла быть в дальнейшем исправлена, например, другим экспертом интегрированной ИС и эти изменения автоматически были бы восприняты интегрированной ИС. Итак, необходимо разработать механизм, максимально автоматизирующий разрешение семантических конфликтов, который бы, с одной стороны, минимизировал вмешательство экспертов в предметной области, самостоятельно разрешая конфликты, а с другой, позволял бы экспертам гибко вмешиваться и корректировать результаты семантического сопоставления сущностей.

Следует отметить, что для высокой эффективности взаимодействия предметного посредника с оболочками ИС необходимо все сущности предметной области представлять с помощью целочисленных идентифи-

каторов. То есть необходимо обеспечить централизованную сквозную нумерацию всех сущностей предметной области в рамках интегрированной ИС. Это позволит оболочкам ИС использовать единые идентификаторы сущностей для взаимодействия с предметным посредником, что повысит эффективность обработки запросов. Таким образом, необходимо предусмотреть централизованный механизм сквозной нумерации сущностей предметной области по запросу оболочек интегрируемых ИС и передачу оболочкам присвоенных идентификаторов, которые должны ими использоваться для ответа на запросы предметного посредника. Для разрешения семантических конфликтов в интегрированной ИС предлагается следующая схема (рис. 3.4.3).

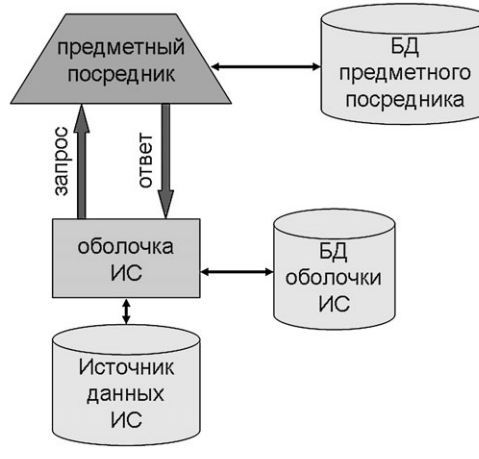


Рис. 3.4.3. Схема разрешения семантических конфликтов в интегрированной ИС

Основная идея заключается в том, что для получения уникальных идентификаторов, описывающих сущности предметной области, оболочка интегрированной системы должна посылать запросы предметному посреднику. До получения уникальных идентификаторов от предметного посредника оболочка не имеет права ссылаться при ответе на запросы предметного посредника на объекты интегрируемой ИС, глобальные идентификаторы которых неизвестны. Это обеспечит правильную работу.

Рассмотрим эту схему более подробно на примере. Допустим, интегрируемая ИС содержит информацию по свойствам химических систем, веществ и их модификаций. Для того чтобы оболочке ИС иметь право

ссылаться на химические системы, описываемые в ИС, ей необходимо сформировать XML-документ, содержащий список описаний химических систем, глобальные идентификаторы которых неизвестны. При этом химическая система s должна быть представлена, как множество химических элементов e_i: $s = \{e_1, e_2, \ldots, e_n\}$. После формирования XML-документа, оболочка ИС посылает запрос Web-сервису предметного посредника. Предметный посредник анализирует каждую химическую систему $s = \{e_1, e_2, \ldots, e_n\}$ из XML-документа и пытается отыскать описание этой системы в своей БД, формируя при этом XML-документ, содержащий ответ для оболочки ИС. Если предметный посредник обнаруживает химическую систему в своей БД, то добавляет в ответ глобальный идентификатор химической системы, найденный в БД. Если же предметному посреднику не удается найти в БД химическую систему, соответствующую множеству заданных элементов, то предметный посредник добавляет в свою БД описание новой химической системы и присваивает ей уникальный глобальный идентификатор, добавляя его в документ-ответ.

Аналогично происходит определение глобальных идентификаторов химических веществ. Как и в химической системе, помимо множества химических элементов, вещество c представляется функцией f, задающей количественное вхождение каждого e_i химического элемента в веществе c. В БД предметного посредника для веществ, так же, как и для химических систем, ведется справочник, содержащий глобальные идентификаторы веществ. Химическое вещество, идентификатор которого запрашивается у предметного посредника, считается эквивалентным веществу, данные о котором уже содержатся в БД предметного посредника, тогда и только тогда, когда функция f запрашиваемого вещества эквивалентна функции f вещества, информация о котором содержится в БД (с учетом соответствующих множеств химических элементов e_i).

Следует отметить, что описания химических систем и веществ довольно хорошо формализованы. В периодической системе Д. И. Менделеева приведены обозначения всех химических элементов, которые могут присутствовать в системах и веществах. Есть соглашения по записи химических формул, которые позволяют разработать программные модули, выполняющие автоматическое преобразование химических формул, например, из HTML-описания в XML-документ нужного формата, который и используется для обмена информацией между узлами ИС.

Сложнее обстоит вопрос с идентификацией кристаллических модификаций, обозначение которых не стандартизировано в разных ИС СНВМ. В результате возникают трудноразрешимые семантические конфликты при обозначении одних и тех же модификаций различными строковыми литера-

лами. Схожая ситуация наблюдается и при обозначении свойств. Так при описании одного и того свойства могут использоваться несколько отличающиеся термины, которые вычислительной системой не могут восприниматься как эквивалентные, хотя по своей семантике таковыми и являются.

Для разрешения сложных семантических конфликтов, которые возникают при описании кристаллических модификаций, в интегрированной ИС предлагается использовать несколько более сложную модель разрешения конфликтов, чем та, что описана выше. Суть этой модели заключается в том, что предметный посредник наряду с глобальным идентификатором сущности, запрашиваемым оболочками ИС, должен также передавать статус этого глобального идентификатора. Роль статуса, фактически ассоциированного с каждым глобальным идентификатором (и, соответственно, с каждой сущностью), заключается в том, что он показывает, насколько достоверным данный глобальный идентификатор является, т. е. может ли он измениться для указанной сущности в будущем.

Поясним логику работы механизма разрешения семантических конфликтов на примере выявления глобальных идентификаторов кристаллических модификаций. При старте интегрированной ИС все справочники, связанные с глобальными идентификаторами известных сущностей предметной области, в БД предметного посредника являются пустыми. Таким образом, если предметному посреднику придет запрос на получение глобального идентификатора кристаллической модификации, обозначаемой строковым литералом «триклинная», предметный посредник, не найдя ее в справочнике модификаций, будет вынужден добавить ее туда. При этом этой модификации присваивается уникальный глобальный идентификатор, например, 1 и статус этого идентификатора устанавливается в состояние «ненадежный». Это состояние статуса будет обозначать возможную смену глобального идентификатора в будущем. Таким образом, передавая глобальный идентификатор, предметный посредник оповещает оболочку ИС о том, что идентификатор модификации «триклинная» может измениться, и оболочке ИС необходимо через некоторое время вновь запросить глобальный идентификатор со статусом надежности для данной модификации.

Изменить статус идентификатора может только эксперт в предметной области. То есть, увидев, что в БД предметного посредника присутствует только одна запись для триклинной модификации, он может выставить статус для идентификатора в состояние «надежный». Таким образом, оболочка ИС, запросив в следующий раз глобальный идентификатор для модификации «триклинная», получит идентификатор 1 и статус «надежный», что избавит оболочку от необходимости периодически вновь узнавать глобальный идентификатор для модификации «триклинная».

Теперь представим, что оболочка другой интегрируемой ИС запрашивает идентификатор для модификации «тр2. Предметный посредник,

естественно, неудачно сравнив этот литерал с литералом «триклинная», добавит новую запись в справочник модификаций и вернет, например, идентификатор 2 со статусом «ненадежный». Затем, через некоторое время, эксперт в предметной области увидит, что появилась новая модификация «тр» со статусом «ненадежный». Он сопоставит эту запись с уже имеющимися и обнаружит, что это всего лишь другой вариант обозначения сущности «триклинная». При этом он добавит в список синонимов для модификации «триклинная» литерал «тр», а для самостоятельной записи «тр» выставит статус в «удален», который обозначает, что данный глобальный идентификатор, в нашем случае, 2, теперь не используется. При следующей итерации оболочка ИС попытается выяснить глобальный идентификатор модификации «тр», так как она получила в прошлый раз статус идентификатора «ненадежный». Предметный посредник в ответ на этот запрос вернет идентификатор 1 (соответствующий модификации «триклинная») со статусом «надежный».

Следует отметить, что запись для модификации «тр» в справочнике модификаций БД предметного посредника со статусом «удален» фактически становится ненужной, т. к. литерал «тр» попадает в тезаурус синонимов модификации «триклинная». Такие записи со статусами «удален» могут удаляться из БД предметного посредника, например, через месяц после выставления соответствующего статуса, чтобы не накапливать ненужные записи и, тем самым, не засорять БД.

Аналогичный механизм предлагается использовать и при разрешении семантических конфликтов, связанных с названиями описываемых в интегрируемых ИС свойств неорганических веществ.

Итак, подводя итоги, следует отметить, что предложен комплексный подход, который направлен на решение трех основных конфликтов гетерогенности в исследуемой предметной области. Таким образом, применение принципов разрешения конфликтов, описанных в этой главе, позволит построить ИС по свойствам неорганических веществ, интегрированную на уровне источников данных. Организованная таким образом единая ИС будет объединять информацию из всех информационных источников на основе подхода Local-as-View.

Для объединения расчетных подпрограмм и существующих Web-интерфейсов интегрируемых ИС, был предложен метод интеграции EAI. Параллельное применение методов EAI, EII и ETL позволит построить не имеющую аналогов интегрированную ИС СНВМ. Данная ИС сможет широко использоваться не только специалистами в области химии, но и системами поддержки принятия решений (СППР) для анализа данных в ИС, нахождения взаимосвязей и построения гипотез о существовании веществ с заданными характеристиками.

3.5. Платформа для разработки интегрированной ИС СНВМ

Разработка информационной системы с «нуля» может представляться оправданной лишь в том случае, если имеется ярко выраженная специфика предметной области, и применение типовых программных решений попросту неприемлемо. Следует отметить, что это достаточно редкое явление. Большинство современных информационных систем используют в своей основе современное программное обеспечение. К нему относятся операционные системы, системы управления базами данных и другое программное обеспечение, использование которого целесообразно при реализации конкретных информационных систем.

В настоящее время существует большой выбор программной инфраструктуры для реализации информационных систем. При построении любой информационной системы выбор программной платформы является важным этапом, значение которого трудно переоценить, так как неправильный выбор может значительно затруднить реализацию идей, внедрение и дальнейшую поддержку проекта. И наоборот, удачно выбранная программная платформа значительно упрощает построение информационной системы, ее дальнейшую поддержку и сопровождение.

Правильный выбор технологической платформы приведет к минимизации издержек при создании информационной системы и повышению экономической эффективности не только процесса разработки, но и большинства этапов жизненного цикла информационной системы [121].

Наиболее эффективным представляется подход, при котором за основу построения информационной системы выбирается стандартное программное обеспечение и инструментарий, позволяющий полностью описать информационную систему, при необходимости гибко достраивать и расширять ее функции, а также способный интегрироваться с другими системами, в первую очередь, уже эксплуатирующимися. Последнее особенно важно, учитывая, что базы данных по свойствам веществ для электронной техники разрабатывались на различных аппаратных и программных средах.

Очевидно, что в настоящее время рынок программного обеспечения достаточно богат, и на нем есть группа крупных поставщиков, предлагающих схожие по функциональным возможностям программные продукты. Среди таких компаний можно выделить Microsoft, IBM, Sun Microsystems, Oracle, BEA, Computer Associates и др. Для определения программной платформы, на базе которой будет осуществляться построение информационной системы нужно провести всесторонний анализ продуктов ведущих поставщиков. При этом следует отметить, что для повышения объективности проводимого анализа необходимо использовать результаты

сравнений в данной области, проведенных крупными и авторитетными независимыми аналитическими компаниями.

Рассмотрим критерии, которые будут использованы для выбора технологической платформы. Здесь необходимо отметить, что все критерии можно условно разделить на три группы:

- *Функциональные* — эта группа критериев должна отражать, насколько полно платформы обеспечивают все требования, предъявляемые к функциональным возможностям создаваемой информационной системы. Следует также учитывать полноту поддержки созданной модели данных в рамках выбранной технологической платформы.
- *Экономические* — эта группа критериев оценивает экономическую эффективность предлагаемых технологических платформ (стоимость общего и специализированного программного обеспечения, расходы на внедрение и сопровождение системы).
- *Технологические* — эта группа критериев отражает насколько полно платформа поддерживает общепринятые стандарты при разработке информационных систем и возможность интеграции с уже эксплуатируемым программным обеспечением. Эти критерии включают также требования к аппаратным ресурсам.

Что касается группы функциональных критериев, то все современные программные платформы обладают достаточной гибкостью для решения различных классов задач. В связи с этим требуется исследовать возможности программных платформ по экономическим и технологическим критериям. Критерии выбора программной платформы предлагаются следующие:

- производительность;
- безопасность;
- надежность;
- интероперабельность;
- совокупная стоимость владения (ССВ).

3.5.1. Производительность

В связи с тем, что информационная система будет построена на базе Web-сервера, наиболее уместным будет сравнение наиболее популярных Web-серверов, использующихся на сегодняшний день:

- Web-сервер Internet Information Server (IIS) на платформе Microsoft Windows Server 2008;
- Web-сервер Apache на базе высокопроизводительных коммерческих Unix-платформ;

- Web-сервер Nginx на базе высокопроизводительных коммерческих Unix-платформ;
- Web-сервер Lighttpd на базе высокопроизводительных коммерческих Unix-платформ;
- Web-сервер G-WAN на базе высокопроизводительных коммерческих Unix-платформ.

Для сравнения используются результаты испытаний опубликованные на сайте компании Web Performance, Inc в ноябре 2011 [122]. В этих тестах сравнивалась пиковая производительность по числу обрабатываемых запросов в секунду, показываемая указанными выше Web-серверами на одном и том же аппаратном обеспечении. Основной вывод по результатам теста, гласит, что Web-сервер IIS 7.0 на базе Windows Server 2008 позволяет обслуживать большее число одновременно подключенных пользователей, чем Apache 2.x, Nginx, Lighttpd и G-WAN на базе Red Hat Enterprise Linux 6.0.

При этом на тестах статических Web-страниц преимущество платформы Microsoft над Unix-платформами составляло от 2 % на Web-сервере G-WAN и до 350 % на Web-сервере Apache 2.x (рис. 3.5.1).

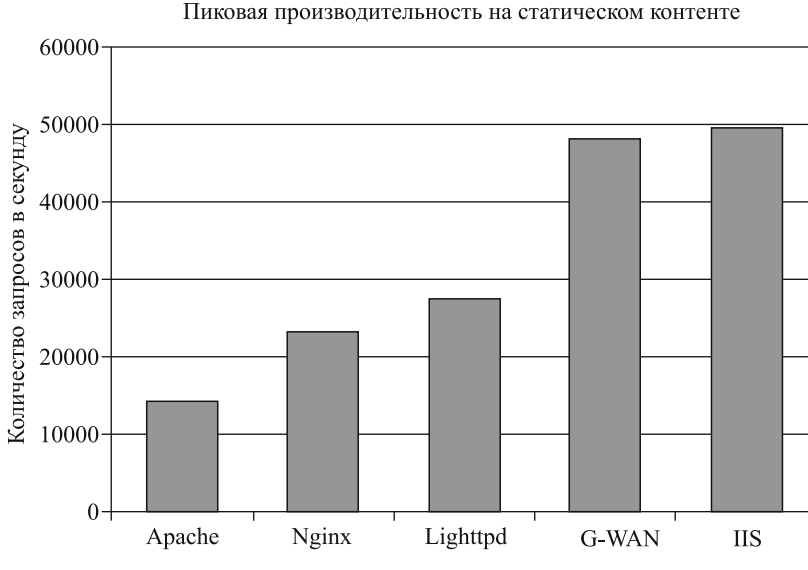


Рис. 3.5.1. Сравнение производительности на статических Web-страницах [122]

На тестах также была оценена нагрузка на центральный процессор. Меньшее значение подразумевает снижение затрат на оборудование и электроэнергию, а также представляет избыток производственных мощностей, что предположительно может быть использовано для выполнения других задач (рис. 3.5.2).

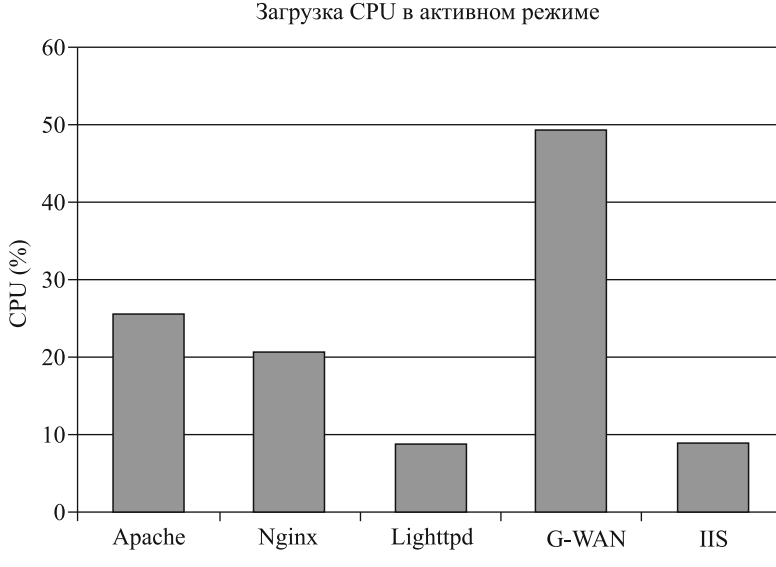


Рис. 3.5.2. Сравнение нагрузки на CPU Web-серверами [122]

Следует отметить, что Web-сервер IIS 7.0 (рис. 3.5.2) на Windows-платформе значительно меньше загружает процессор при активной работе (9 %), чего не скажешь об его ближайших конкурентах Apache 2.x на Unix-платформе (26 %) и Nginx на Unix-платформе (21 %). Web-сервер G-WAN на Unix-платформе хоть и показал хорошие результаты производительности, но нагрузка на процессор при работе составляет 50 %.

Следует отметить, что вышеприведенная диаграмма (рис. 3.5.2) не отражает реальный потенциал Windows-платформы, поскольку при сравнении производительности для формирования динамического Web-содержимого использован интерфейс CGI. В настоящее время для информационных систем, требующих высокой производительности, Microsoft рекомендует применять технологию ISAPI, потенциал которой раскрывает следующая диаграмма (рис. 3.5.3).

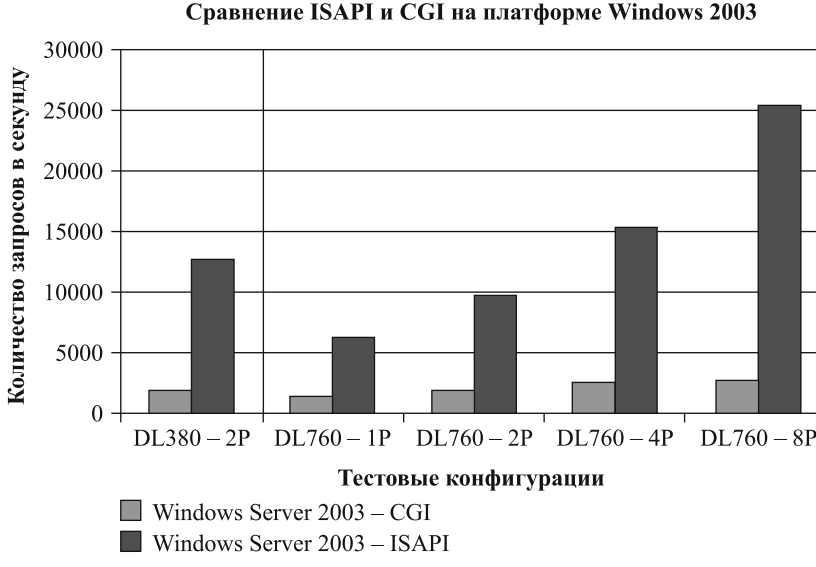


Рис. 3.5.3. Сравнение производительности ISAPI и CGI на платформе Windows 2003 + IIS 6.0 [123]

Как видно из результатов сравнения проведенного Veritest, платформа Windows 2008 Server + IIS 7.0 является наиболее производительной при построении информационных систем на основе Web-приложений.

3.5.2. Безопасность

В настоящее время вопросам информационной безопасности неслучайно уделяется достаточно пристальное внимание. Чем более безопасна система, тем ниже вероятность того, что злоумышленник сможет получить доступ к ценной информации или приостановить работу системы. В феврале 2010 года Кристиан Флориан — сотрудник компании GFI Software, опубликовал доклад, в котором проводится обзор наиболее уязвимых операционных систем, предоставляемых ведущими производителями. В докладе рассматриваются информационные уязвимости, выявленные в период с января 2010 года по декабрь 2010, и степень их опасности для системы [124].

Одним из важнейших параметров при оценке безопасности платформы является среднее число дней, которое потребовалось разработчику ПО с момента обнаружения уязвимости в платформе до момента предоставления программной заплатки (security patch), устраняющей уязвимость.

На диаграмме показано среднее число дней, требующееся разработчику для полного устранения выявленных уязвимостей (рис. 3.5.4).

Важным компонентом безопасности информационной системы является также безопасность кода, который лежит в ее основе. Чем меньше уязвимостей было выявлено, тем более безопасным является код, и тем выше безопасность информационной системы. Общее число уязвимостей в продуктах основных поставщиков операционных приведено в таблице (табл. 3.3).

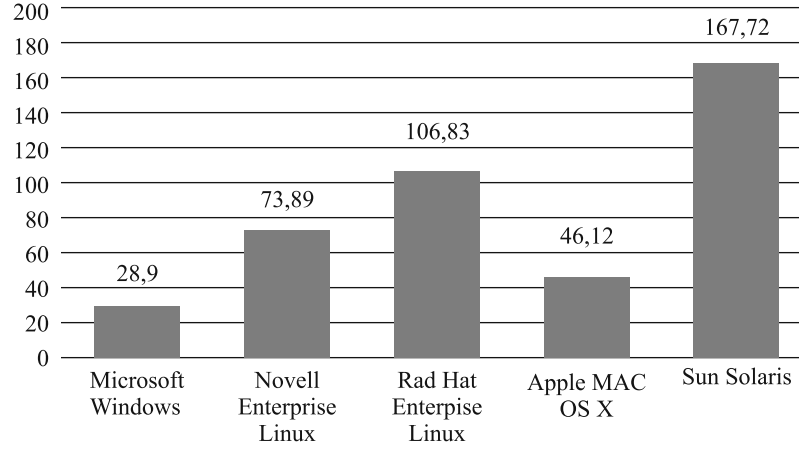


Рис. 3.5.4. Среднее время (в днях), требующееся на устранение уязвимостей платформ ведущих поставщиков ПО [125]

Таблица 3.3. Общее число и степень опасности найденных уязвимостей [124]

ОС	Степень опасности			
	Сумма	**Высокая**	**Средняя**	**Низкая**
Microsoft Windows Server 2003	147	97	50	0
Microsoft Windows XP	100	70	30	0
Microsoft Windows Vista	88	64	24	0
Microsoft Windows Server 2008	92	61	31	0
Microsoft Windows 7	66	50	16	0
Linux Kernel	129	33	58	38
Apple Mac OS X Server	103	22	76	5
Apple Mac OS X	96	20	72	4
Cisco IOS	28	26	2	0

3.5.3. Надежность

Под надежностью программной платформы обычно понимается ее отказоустойчивость, т. е. время вынужденного простоя и оперативность решения производителем программного обеспечения проблем, вызванных ошибками в коде операционной системы [126].

Сокращения количества и среднего времени простоев можно добиться с помощью:

- непрерывного мониторинга состояния ИС;
- оперативного оповещения специалистов о сбоях;
- автоматического выполнения корректирующих действий;
- удобных средств диагностики;
- раннего обнаружения условий, предшествующих сбою;
- раннего обнаружения нехватки ресурсов.

Что касается времени вынужденного простоя систем на базе Windows и Unix, исходя из графиков среднего времени нахождения и устранения уязвимостей системы, то Unix явно уступает Microsoft по времени оперативного решения проблем, связанных с «дырами» в системе. Также присутствуют данные по простою платформ предыдущего поколения [127].

Согласно этому исследованию, системы на базе Microsoft Windows Server 2003 имели 4 часа 20 минут 19 секунд простоя, в то время как системы на базе Red Hat Enterprise Linux AS 3.0 имели 4 часа 59 минут 44 секунды простоя при выполнении тестовых задач (рис. 3.5.5).

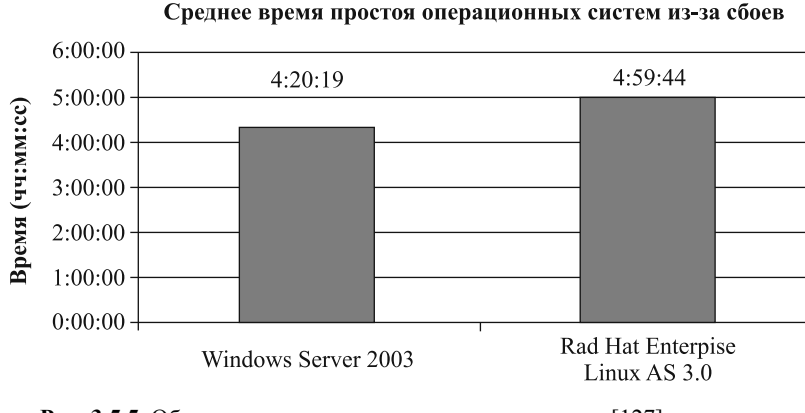


Рис. 3.5.5. Общее время простоя операционных систем [127]

3.5.4. Интероперабельность

Под интероперабельностью (interoperability — способность к взаимодействию с другими ИС) понимается способность программной среды взаимодействовать с разнородными программными средами, применяемыми на других программно-аппаратных платформах [128]. В настоящее время в качестве общепринятого стандарта такого взаимодействия являются Web-сервисы (Web Services) как основа сервисно-ориентированной архитектуры (Service Oriented Architecture), активно продвигаемой консорциумом W3C [129].

В настоящий момент заказчики ИТ-решений требуют единую и хорошо управляемую ИТ-инфраструктуру вне зависимости от того, используются ли продукты только одного производителя или нескольких. Microsoft и сообщество разработчиков GNU/Linux проделали большую работу, чтобы интегрировать Linux и ключевые серверные продукты Microsoft между собой. Более того, в последние годы Microsoft сам стал одним из заметных разработчиков ядра Linux, делая акцент на вопросах интероперабельности.

Microsoft и Linux центр запустили долгосрочную программу по разработке технической документации, которые позволят администраторам, работающим в смешанных GNU/Linux-Windows сетях выстраивать единую и хорошо управляемую систему. В качестве исполнителя был выбран Национальный исследовательский университет «МЭИ», в котором с 2008 года функционирует Центр инноваций Microsoft, а также набрана большая экспертиза по GNU/Linux и Свободному ПО.

Согласно исследованию «Interoperability: How Technology Managers Rate Microsoft and Its Technologies for Development», проведенному фирмой Jupiter Research [130], 72 % опрошенных специалистов отдали свои голоса компании Microsoft, высоко оценив интероперабельность, обеспечиваемую технологическими решениями этой компании. В то же время 55 % респондентов отметили основанный на XML протокол SOAP и соответствующий ему язык WSDL, как наиболее ценные технологии для обеспечения интероперабельности (сам язык XML с 37 % предпочтений респондентов занимает почетное второе место). Это исследование подтверждает огромную роль SOA при интеграции гетерогенных информационных систем.

Компания ObjectWatch провела исследование возможности взаимодействия информационных сред с использованием архитектуры, ориентированной на сервисы (SOA) [131]. В этом исследовании рассматриваются технологии .Net, J2EE, CORBA, различные операционные системы и их способность реализовать SOA-архитектуру. Технологии исследуются с различных точек зрения:

- поддержка открытых стандартов;
- совокупная стоимость владения (ССВ);

- продуктивность инструментальных средств;
- безопасность;
- масштабируемость;
- производительность;
- надежность.

Следует отметить, что за компанией Microsoft признается роль лидера в архитектурах, ориентированных на сервисы. Особо подчеркивается, что это достигается не только благодаря прекрасной поддержке Web-сервисов, но и за счет поддержки интероперабельности в серверных продуктах, таких как SQL Server, BizTalk Server и Host Integration Server.

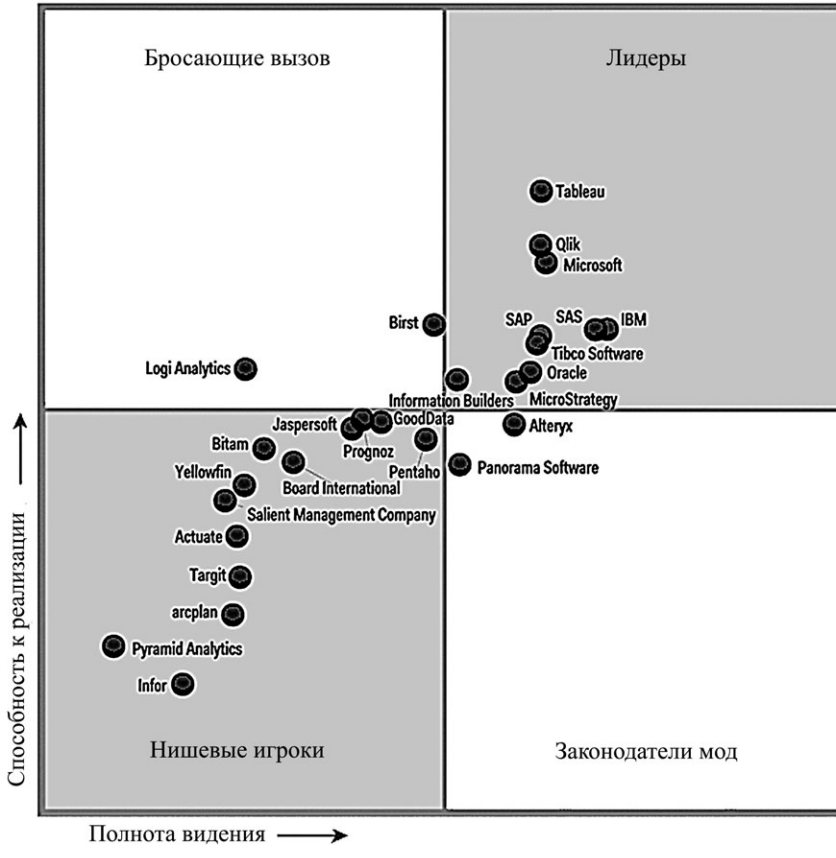


Рис. 3.5.6. «Волшебный квадрат» влияния основных поставщиков Web-сервисов, полученный Gartner Group в феврале 2014 года [132]

Если же рассматривать только Web-сервисы, то и здесь позиции Microsoft незыблемы. Эта компания играет роль законодателя мод и совместно с W3C и другими независимыми организациями, работающими над открытыми технологиями, участвует в разработке и успешно реализует стандарты в области Web-сервисов. В качестве подтверждения этих слов можно привести «волшебный квадрат», полученный Gartner Group [132] при исследовании влиятельности основных поставщиков программного обеспечения (рис. 3.5.6).

3.5.5. Совокупная стоимость владения

Совокупная стоимость владения (ССВ, TCO — Total Cost of Ownership) — это интегральный показатель экономической эффективности применяемых технологий. Существует несколько методик вычисления ССВ, разработанных ведущими аналитическими компаниями (Gartner Group, Forrester Research, Meta Group, IDC), но все они нацелены на определение наилучшего соотношения цена/качество для оборудования и ПО. В рамках этих подходов предполагается оценка стоимости приобретения, администрирования, установки, перемещения и модернизации, технической поддержки и сопровождения, вынужденных простоев и других скрытых затрат. В этот показатель включается как стоимость лицензий на используемое программное обеспечение, так и затраты на его дальнейшую поддержку и сопровождение.

Приведем результаты исследования совокупной стоимости владения Microsoft Windows и Linux, представленные сотрудником компании ZDNet в августе 2008 года (рис. 3.5.7) [133].

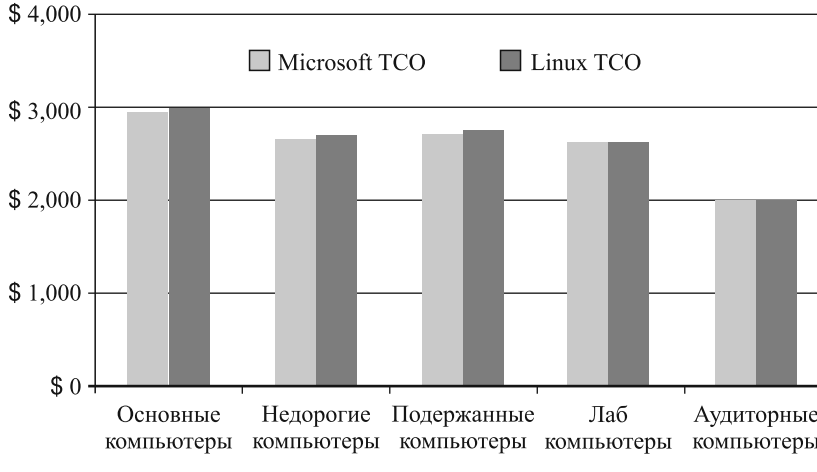


Рис. 3.5.7. Исследования ССВ Microsoft Windows в сравнении с ССВ Linux [133]

Таблица 3.4. Десятка лучших систем в тестах TPC-E по производительности [46]

Rank	Company	System	Performance (tpsE)	Price/tpsE	Watts/tpsE	System Availability	Database	Operating System	Processors / Cores / Threads	Date Submitted
1	IBM	IBM System x3850 X6	5,576.27	188.69 USD	NR	04/15/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 Standard Edition	4 / 60 / 120	02/16/14
2	IBM	IBM System x3850 X5	5,457.20	249.58 USD	NR	03/08/13	Microsoft SQL Server 2012 Enterprise Edition	Microsoft Windows Server 2012 Standard Edition	8 / 80 / 160	03/08/13
3	NEC	NEC Express5800/A2040b	5,087.17	229.04 USD	NR	04/15/14	Microsoft SQL Server 2014 Enterprise Edition	Microsoft Windows Server 2012 Standard Edition	4 / 60 / 120	02/17/14
4	NEC	NEC Express5800/A1080a-E	4,614.22	450.18 USD	NR	04/02/12	Microsoft SQL Server 2012 Enterprise Edition	Microsoft Windows Server 2008 R2 Enterprise Edition SP1	8 / 80 / 160	03/27/12
5	IBM	IBM System x3850 X5	4,593.17	140.56 USD	NR	08/26/11	Microsoft SQL Server 2008 Enterprise Edition R2	Microsoft Windows Server 2008 R2 Enterprise Edition SP1	8 / 80 / 160	08/26/11
6	FUJITSU	PRIMERGY RX900 S2	4,555.54	217.27 USD	1.00	07/01/11	Microsoft SQL Server 2008 Datacenter Edition R2	Microsoft Windows Server 2008 R2 Enterprise Edition SP1	8 / 80 / 160	06/06/11
7	FUJITSU	PRIMEQUEST 1800E2	4,414.79	226.19 USD	1.09	07/01/11	Microsoft SQL Server 2008 Enterprise Edition R2	Microsoft Windows Server 2008 R2 Enterprise Edition SP1	8 / 80 / 160	07/27/11
8	NEC	NEC Express5800/A1080a-E	4,200.61	287.42 USD	NR	08/31/11	Microsoft SQL Server 2008 Enterprise Edition R2	Microsoft Windows Server 2008 R2 Enterprise Edition SP1	8 / 80 / 160	04/28/11
9	IBM	IBM System x3850 X5	3,218.46	225.30 USD	NR	11/28/12	Microsoft SQL Server 2012 Enterprise Edition	Microsoft Windows Server 2012 Standard Edition	4 / 40 / 80	11/28/12
10	HUAWEI	Huawei Tecal RH5885 V2	3,053.84	352.48 USD	NR	10/30/12	Microsoft SQL Server 2012 Enterprise Edition	Microsoft Windows Server 2008 R2 Enterprise Edition SP1	4 / 40 / 80	12/14/12

Совокупная стоимость владения Microsoft ненамного, но все же дешевле обходится для компаний. В связи с более высокой производительностью и меньшем потреблением ресурсов процессора в сравнении с Unix-системами [133, 134], то Microsoft заметно впереди своих конкурентов.

При сравнении мощных промышленных СУБД (Microsoft SQL Server, IBM DB2 и Oracle 10g) в тестах TPC-E по производительности платформа на базе Microsoft SQL Server и Windows — явный лидер (табл. 3.4). Данные по состоянию на февраль 2014 года взяты с Web-сайта Transaction Processing Council [46] — независимой некоммерческой организации, измеряющей производительность СУБД по установленным методикам.

Выводы по выбору платформы

Учитывая результаты тестирования, проведенные независимыми компаниями по выбранным нами критериям, следует сделать вывод, что Microsoft является лидером индустрии программного обеспечения, предлагающим надежные и высокопроизводительные системы. При этом совокупная стоимость владения (ССВ) предлагаемых решений также оказывается ниже, чем у конкурирующих компаний, что и обуславливает выбор решений на платформе Microsoft.

Краткие выводы

В главе получены следующие результаты:

- Проведен системный анализ методов интеграции гетерогенных ИС (EII, ETL, EAI).
- Предложены рекомендации по выбору предпочтительного метода интеграции гетерогенных ИС в зависимости от требований, предъявляемых к результирующей интегрированной ИС.
- Предложена методология интеграции ИС СНВМ, объединяющая преимущества современных методов интеграции ИС.
- Разработана методика консолидации данных по свойствам неорганических веществ с использованием методов хранилищ данных и виртуальной интеграции.
- Рассмотрены конфликты гетерогенности, которые необходимо разрешить при разработке интегрированной информационной системы.
- Разработана архитектура интегрированной ИС СНВМ.
- Проанализированы программные платформы для построения интегрированной ИС СНВМ, выработаны критерии отбора и сделан выбор в пользу платформы Microsoft.

Глава 4
Системный подход к разработке хранилища данных по свойствам неорганических веществ для систем поддержки принятия решений

4.1. Диаграммы потоков данных DFD

В рамках предложенной методологии метод консолидации на основе хранилища данных (ХД) применяется при создании интегрированного источника данных в рамках одной организации. В настоящей главе с использованием средств системного анализа и функционального моделирования ИС создается ХД по свойствам неорганических веществ.

Задачу создания интегрированного ХД можно представить с помощью контекстной диаграммы потоков данных следующим образом (рис. 4.1.1). Данные из исходных информационных источников, которыми являются БД по свойствам неорганических веществ и материалов, должны с использованием ETL-инструментов попадать в ХД.

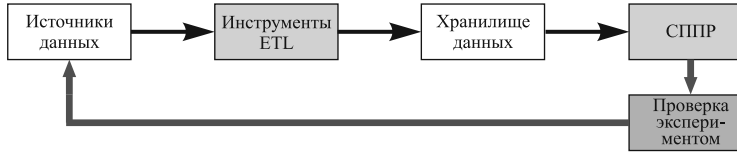


Рис. 4.1.1. DFD-диаграмма хранилища данных совместно с СППР

ХД по свойствам неорганических веществ служит информационной основой для СППР при прогнозировании свойств неорганических веществ. СППР (в роли которой в данной работе выступает информационно-аналитическая система для компьютерного конструирования неорганических соединений) выполняет сложный анализ накопленных данных и на его основе осуществляет прогнозирование. Результаты прогнозов (после

экспертной оценки) должны проходить экспериментальную проверку, направленную на синтез спрогнозированных веществ и оценку их свойств. В этом смысле можно сказать, что прогнозы, полученные с помощью системы компьютерного конструирования, являются важными для существенного сужения области поиска новых функциональных материалов, т. к. СППР позволяет априорно, т. е. до проведения реального химического эксперимента, оценить перспективы синтеза тех или иных веществ.

При использовании в рамках системного подхода декомпозиции блока «инструменты ETL» получим DFD-диаграмму, содержащую модули для извлечения, преобразования и загрузки данных в ХД (рис. 4.1.2).

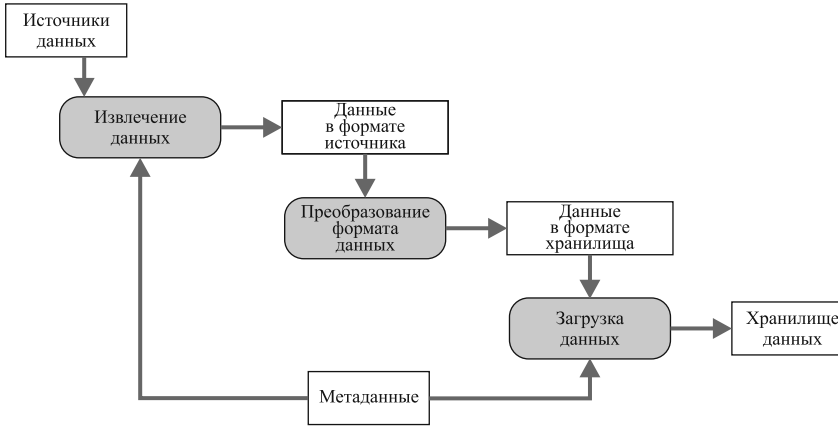


Рис. 4.1.2. DFD-диаграмма декомпозиции блока «инструменты ETL»

4.2. Методология функционального моделирования IDEF0

Для более полного описания информационных процессов использована методология функционального моделирования IDEF0. Контекстная диаграмма функциональной модели методов интеграции данных представлена на рис. 4.2.1. Отметим, что эта диаграмма одинакова для всех методов интеграции данных (EII и ETL), т. к. отличия проявляются на более позднем этапе декомпозиции функциональной модели методов интеграции.

Первый уровень декомпозиции представлен на рис. 4.2.2. В блоке A1 происходит извлечение данных из ИС источников данных на основе метаданных о правилах и способах извлечения информации. Блок A2 с помощью

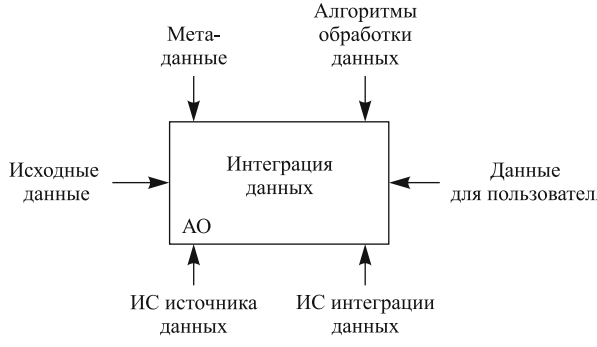


Рис. 4.2.1. Контекстная диаграмма функциональной модели методов интеграции данных

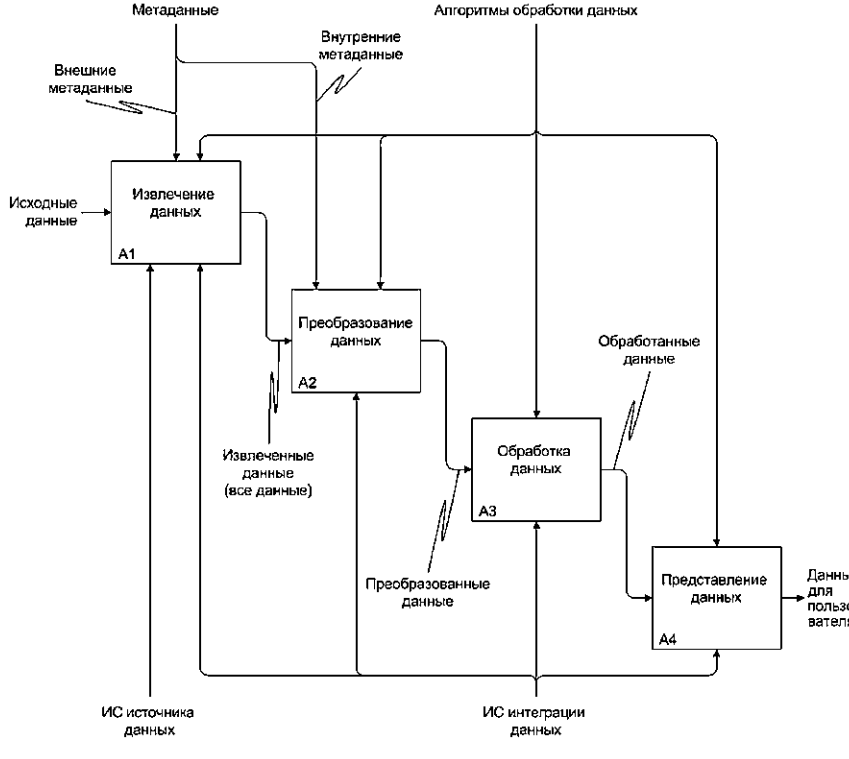


Рис. 4.2.2. Первый уровень декомпозиции функциональной модели методов интеграции данных

специальных алгоритмов осуществляет преобразование извлеченных данных из форматов ИС источников данных к форматам данных, принятых в качестве стандартов для хранилища данных. Блок А3, получая на входе преобразованные к единому формату данные, осуществляет их обработку в соответствии с требованиями интегрированной ИС. Блок А4 использует алгоритмы для поиска затребованных пользователем данных и их представления в удобной форме.

4.3. ER-модель хранилища данных

На основе предложенной и иерархии химический понятий (система → вещество → модификация) была разработана ER-модель ХД (рис. 4.3.1). Следует учесть, что связи многие ко многим (N : M) потребуют ввода дополнительных отношений на этапе реализации реляционной структуры ХД [135].

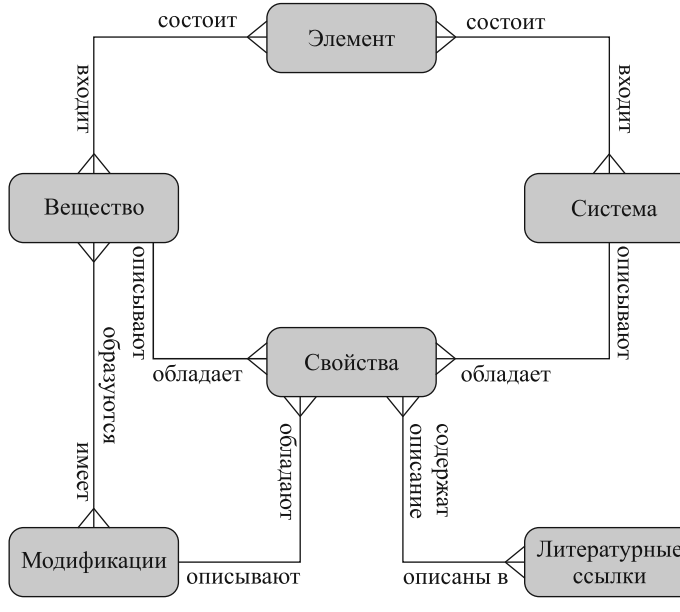


Рис. 4.3.1. ER-модель ХД

4.4. Реляционная структура ХД

В главе 3 было принято решение использовать продукты компании Microsoft для реализации ХД. Для управления ХД, лежащим в основе интегрированной ИС на уровне организации, было принято решение использовать Microsoft SQL Server 2008. Структура ХД базируется на описанной выше ER-модели. На рис. 4.4.1 приведена часть логической модели ХД, включающая все сущности с ER-диаграммы, кроме литературных ссылок.

Рассмотрим подробнее назначение таблиц полученной схемы данных. В таблице DW_Systems хранится информация о химических системах, данные о которых помещены в хранилище. При этом информация об образующих систему химических элементах содержится в таблице DW_SystemElements, ссылающейся на справочник химических элементов DW_Elements. В поле DW_Systems.SystemXML содержится XML-документ, описывающий химическую систему. Формат XML-документов будет рассмотрен ниже (раздел 4.5.3) при описании процедуры прообразования данных. Аналогично описывается информация о содержащейся в ХД информации по химическим соединениям с помощью таблицы DW_Compounds, связанной с таблицей элементов через промежуточную таблицу DW_CompoundElements, позволяющую хранить сведения о соотношении химических элементов в соединении (возможно, переменного состава) с помощью полей MinIndex и MaxIndex. Информация о кристаллических модификациях хранится в таблице DW_Modifications, которая добавляет к данным о количественном составе соединения информацию о типе одной из кристаллических модификаций, описанных в справочной таблице DW_ModificationTypes.

Основное внимание уделялось способу представления значений свойств. Важно обеспечить представление свойств разных типов — скалярные значения, табличные наборы данных, графические и полнотекстовые описания. Для обеспечения кроссплатформенной возможности работы со значениями из столь широкого диапазона типов данных было принято решение использовать XML-документы для представления значений свойств. При работе с двоичными данными, например, с графическими рисунками или аналитическими обзорами используется представление бинарных данных в виде строки Base64 с обязательным указанием MIME-типа ресурса (например, «image/gif» или «application/pdf»). Сами значения хранятся в таблице DW_PropertyValues, а составной ключ PropertyID, SystemID, CompoundID, ModificationID указывает на описываемое свойство из таблицы DW_Propeties и химическую сущность, определяемую тройкой (SystemID, CompoundID, ModificationID). Это позволяет сохранять в ХД значения свойств для химических сущностей на уровне систем,

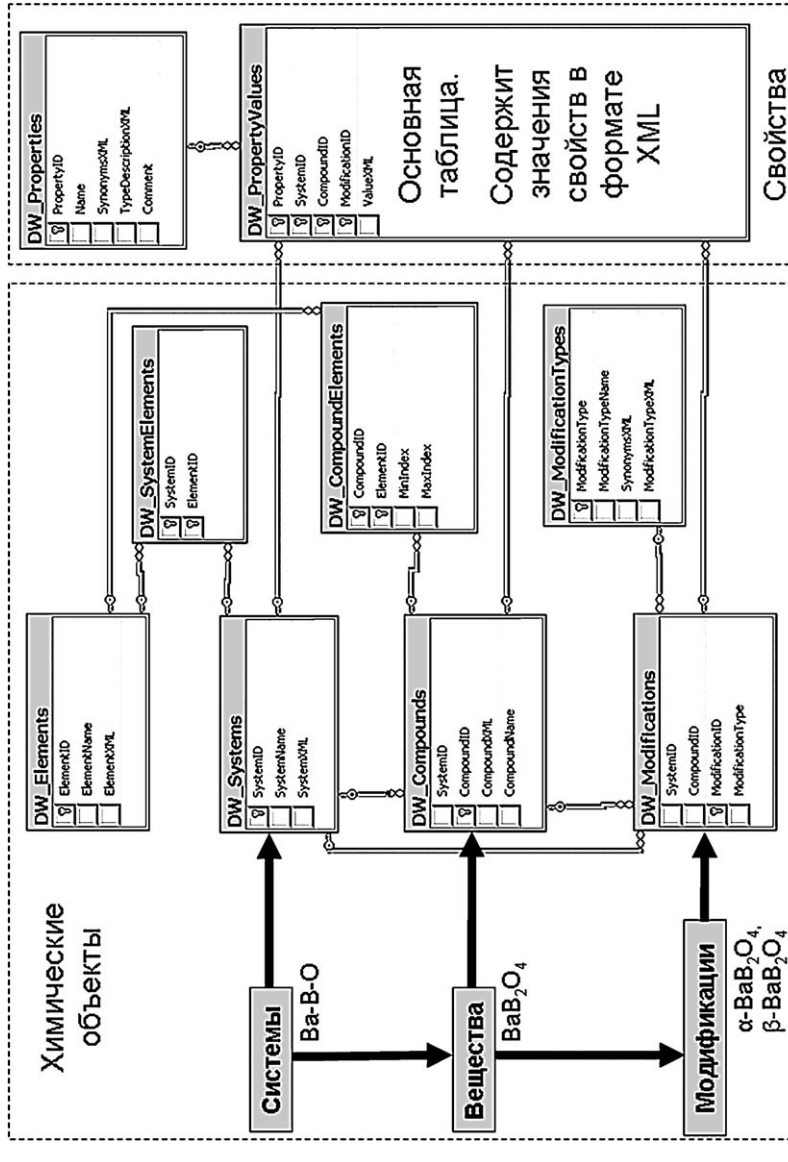


Рис. 4.4.1. Логическая модель ХД по свойствам неорганических веществ

веществ и модификаций. Так, при описании значения свойства для неорганического вещества CompoundID содержит идентификатор соответствующего химического соединения (SystemID — указывает на соответствующую химическую систему), а ModificationID = 0, т. е. указывает на отсутствие информации по модификации химической сущности (использование null недопустимо, в силу вхождения поля ModificationID в состав первичного ключа).

Отмечается, что при использовании ХД в качестве источника информации для систем прогнозирования значение свойства необходимой химической сущности может быть получено из XML-формата двумя способами: 1) путем наложения на XML-документ специального XSLT-преобразования (XML + XSLT = > требуемый формат данных); 2) путем программной обработки XML документа с использованием средств организации запросов к XML-документу на языках XPath и/или XQuery. Важной особенностью является возможность выполнения данных преобразований как на стороне ХД с использованием хранимых процедур на SQL CLR (SQL Common Language Runtime — реализация размещения и запуска управляемого кода на .Net Framework в рамках СУБД Microsoft SQL Server), так и на стороне сервера приложений (например, Microsoft IIS), что улучшает масштабируемость ИС.

4.5. Извлечение, преобразование и загрузка данных в ХД

Разработана методика извлечения, преобразования и загрузки данных в ХД. Данные извлекаются из исходных источников данных, преобразуются в формат ХД и загружаются в ХД. Последовательность этих процессов является основополагающей при заполнении ХД информацией по свойствам неорганических веществ. На рис. 4.5.1. показано, что данные извлекаются из материаловедческих БД, затем преобразуются на ETL-сервере и загружаются в ХД. В процессе извлечения, преобразования и загрузки информации используются метаданные, которые описывают правила извлечения данных из исходных БД, формат исходных данных, а также правила их преобразования к формату единого ХД и последующей загрузки в него.

4.5.1. Процедура извлечения

Процедура извлечения исходных данных из БД информационных источников является первым шагом в подходе к интеграции средствами ETL. Учитывая то, что исходные ИС (из которых извлекается информация)

построены на основе современных реляционных СУБД, предлагается использование SQL-запросов для извлечения требуемых наборов данных.

В качестве примера рассмотрим процедуру извлечения данных о химических соединениях на основе основной таблицы БД ИС «Кристалл» HeadTabl (рис. 4.5.2).

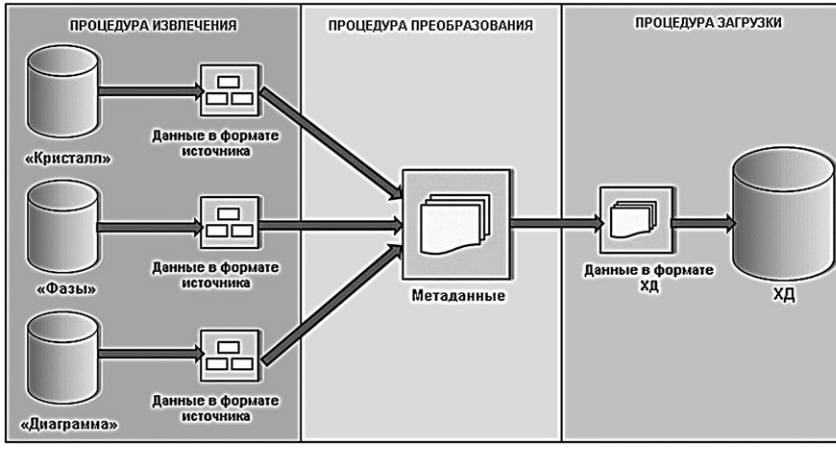


Рис. 4.5.1. Методика извлечения, преобразования и загрузки данных

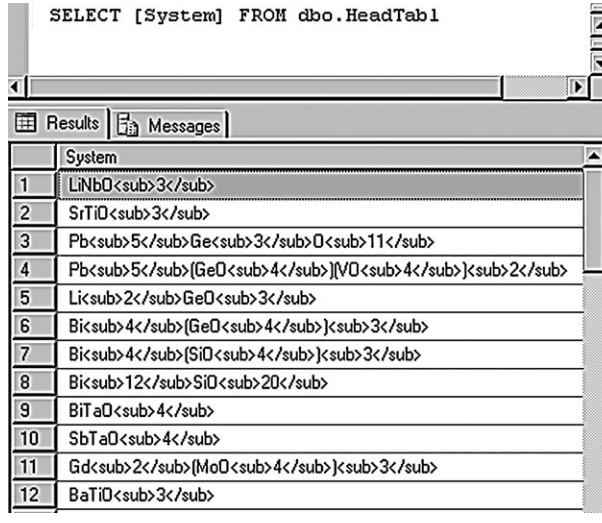
	HeadClue	System	Expert	Help	Class
1	1	LiNbO₃	Буш А.А.	-Li-Nb-O-	0
2	2	SrTiO₃	Буш А.А.	-Sr-Ti-O-	0
3	3	Pb₅Ge₃O₁₁	Буш А.А.	-Pb-Ge-O-	0
4	4	Pb₅(GeO₄)(VO₄)<s...	Буш А.А.	-Pb-Ge-O-V-	0
5	5	Li₂GeO₃	Буш А.А.	-Li-Ge-O-	0
6	6	Bi₄(GeO₄)₃	Буш А.А.	-Li-Ge-O-	0
7	7	Bi₄(SiO₄)₃	Буш А.А.	-Bi-Si-O-	0
8	8	Bi₁₂SiO₂₀	Буш А.А.	-Bi-Si-O-	0
9	9	BiTaO₄	Буш А.А.	-Bi-Ta-O-	0
10	10	SbTaO₄	Буш А.А.	-Sb-Ta-O-	0
11	11	Gd₂(MoO₄)₃	Буш А.А.	-Gd-Ta-O-	0
12	12	BaTiO₃	Буш А.А.	-Ba-Ti-O-	0
13	13	Ba₂NaNb₅O₁₅	Сиротинкин В.П.	-Ba-Na-Nb-O-	0
14	14	Ba_xSr_{1-x}Nb₂O<sub...	Буш А.А.	-Ba-Sr-Nb-O-	0
15	15	K₃Li₂Nb₅O<sub>1...	Буш А.А.	-K-Li-Nb-O-	0
16	16	K(Ta_{1-x}Nb_x)O₃	Буш А.А.	-K-Ta-Nb-O-	0
17	17	KTiOPO₄	Буш А.А.	-K-Ti-O-P-	0
18	18	PbNb₄O₁₁	Буш А.А.	-Pb-Nb-O-	0
19	19	PbZrO₃	Буш А.А.	-Pb-Zr-O-	0
20	20	PbTiO₃	Буш А.А.	-Pb-Ti-O-	0

Рис. 4.5.2. Часть данных таблицы HeadTabl в БД «Кристалл»

Поскольку для наполнения хранилища требуется информация по количественному составу веществ, которая содержится в атрибуте System таблицы HeadTabl, извлекаем ее с помощью соответствующего SQL-запроса:

```
SELECT [System] FROM dbo.HeadTabl
```

После выполнения данного SQL-запроса мы получаем сведения о необходимых химических соединениях, находящихся в данной таблице. Данные извлекаются в формате исходного источника данных ИС. В случае выполнения запроса в среде Microsoft SQL Server Management Studio результат может быть таким, как показано на рис. 4.5.3.



	System
1	LiNbO₃
2	SrTiO₃
3	Pb₅Ge₃O₁₁
4	Pb₅(GeO₄)(VO₄)₂
5	Li₂GeO₃
6	Bi₄(GeO₄)₃
7	Bi₄(SiO₄)₃
8	Bi₁₂SiO₂₀
9	BiTaO₄
10	SbTaO₄
11	Gd₂(MoO₄)₃
12	BaTiO₃

Рис. 4.5.3. SQL-запрос на извлечение данных о химических веществах

По аналогии с данным примером, данные извлекаются не только о химических веществах, но и о кристаллических модификациях и свойствах веществ, содержащихся в БД «Кристалл», «Фазы» и «Диаграмма».

Результатом процедуры извлечения являются данные в табличной форме, которые затем с помощью алгоритмов обработки данных должны преобразовываться в формат хранилища данных для преследующей их загрузки в описанное выше ХД по свойствам неорганических веществ.

4.5.2. Процедура преобразования данных

Гетерогенные ИС СНВМ отличаются форматами представления данных (БД используют различные по синтаксическому описанию и структуре данные). В ряде БД используются реляционные СУБД, в других иерархические СУБД. В последнее время нередко строятся ИС, которые используют XML или какие-либо его известные приложения, например, RDF для хранения информации. В ИС СНВМ, разработка которых велась довольно давно, нередко можно встретить двоичные форматы хранения данных. Все это многообразие форматов данных и способов представления и обработки информации приводит к тому, что БД в том виде, в котором они существуют, зачастую являются несовместимыми с другими программными продуктами. Следует отметить, что изначально при проектировании большинства ИС СНВМ, взаимодействие с внешней программной средой не предусматривалось вовсе.

Разрешить синтаксические и структурные конфликты можно посредством введения общей схемы представления информации и обмена данными, построенной согласно описанию предметной области. При описании химических сущностей, как было показано ранее, можно использовать три уровня: система, вещество и кристаллическая модификация. Следовательно, все подсистемы интегрированного ХД должны оперировать данными типами объектов при ссылке на химические сущности. При этом стоит учитывать, что если описывается определенная химическая модификация, то определена также и химическая система с веществом, модификация которого описывается. То есть если описание химической сущности ведется на уровне модификаций, то все вышележащие уровни (вещество и система) тоже описаны. Следует заметить, что обратное неверно. То есть, если описывается химическая система, то вещество и модификация четко не определены. Однако необходимо понимать, что при описании сущности на уровне системы, все описанные свойства автоматически распространяются на все вещества и модификации, образованные в этой химической системе.

Как было отмечено, в интегрируемых БД содержится информация по различным свойствам химических сущностей. Например, плотность, растворимость, теплопроводность. При этом для каждой химической сущности в БД нередко содержится несколько записей для описания значения свойства. Это обусловлено несколькими обстоятельствами. Во-первых, информация, содержащаяся в БД, может быть взята из различных источников, при этом данные нередко расходятся. Это объясняется как различными способами измерения, так и точностью измеряющей аппаратуры. Таким образом, часто в ИС СНВМ приводится несколько вариантов значения, например, плотности различных кристаллов. Во-вторых, значения

рассматриваемых свойств зачастую зависят от внешних условий, при которых приводились измерения. Например, такие параметры как растворимость и ширина запрещенной зоны зависят от температуры, при которой проводились измерения. Другими словами, значения свойств 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 являться функцией от разного числа аргументов, и поэтому невозможно предложить универсальный формат представления заданного типа свойства для всех ИС СНВМ. 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, которая еще не включена в общее интегрированное ХД, то при ее включении в состав интегрированного ХД возникнет проблема согласования форматов представления указанного свойства. Таким образом, невозможно заранее предусмотреть все зависимости и заложить их в общий формат представления данных даже для отдельно взятого конкретного свойства, не говоря о представлении свойств в целом.

Таким образом, необходим некоторый механизм, позволяющий гибко представлять значения свойств в рамках интегрированного ХД. В настоящее время существует общепризнанное средство описания произвольных форматов данных — это XML (eXtensible Markup Language). С помощью этого языка разметки удобно описывать различные структуры данных, он является межплатформенным форматом и поддерживается большинством языков и библиотек. На сегодняшний день именно этот язык является тем звеном, которое может служить основой для обеспечения взаимодействия различных программно-аппаратных платформ. В настоящее время все большее информации в современных промышленных системах представляется в тех или иных XML форматах.

Формат данных для химических систем

Для разрешения конфликтов разработаны форматы представления описанных химических сущностей и свойств в рамках интегрированного ХД на языке XML. То есть, были разработаны форматы соответствующих XML-документов для представления химических сущностей и их свойств.

Например, химическая система In–S представлена следующим образом:

```
<SystemXML>
    <ChemicalSystem>
        <Item Element = "In" />
<Item Element = "S" />
    </ChemicalSystem>
</SystemXML>
```

Стоит обратить особое внимание на то, что приведенный XML-документ содержит только список химических элементов, входящих в состав системы. Другими словами, задается множество химических элементов {In, S}, принадлежащих конкретной системе.

Формат данных для химических веществ

Описание вещества по сравнению с химической системой дополнительно включает данные о количественном вхождении химических элементов. Рассмотрим пример представления двух разных веществ, относящихся к системе In–S (на самом деле, для системы InS их больше). Так формула вещества состава In_2S_3 будет представлена в XML-формате следующим образом:

```
<SubstanceXML>
<ChemicalSubstance>
        <Item Element = "In" value = "2" />
        <Item Element = "S" value = "3" />
</ChemicalSubstance>
</SubstanceXML>
```

Как видно из XML-документа, количественные вхождения элементов в состав вещества указаны в атрибуте value, принадлежащем соответствующему узлу, описывающему химических элемент. Другим примером вещества в системе In–S является InS (в вещество входит по одному атому индия и серы), которое будет представлено следующим образом:

```
<SubstanceXML>
<ChemicalSubstance>
        <Item Element = "In" value = "1" />
        <Item Element = "S" value = "1" />
</ChemicalSubstance>
</SubstanceXML>
```

Формат данных для кристаллических модификаций

Кристаллическая модификация — это вспомогательный геометрический образ, вводимый для анализа строения кристалла. Поэтому на уровне кристаллических модификаций, помимо информации о количественном составе вещества, содержатся данные о его сингонии (иногда о кристаллической системе). Всего известно шесть сингоний: триклинная, моноклинная, ромбическая, тетрагональная, гексагональная и кубическая.

Разбиение на кристаллические системы выполняется в зависимости от набора элементов симметрии, описывающих кристалл. Такое деление приводит к семи кристаллическим системам, две из которых (тригональная и гексагональная) имеют одинаковую по форме элементарную ячейку и поэтому относятся к одной гексагональной сингонии. Говорят, что гексагональная сингония подразделяется на две подсингонии или гипосингонии. Это обстоятельство учитывается при разрешении конфликтов гетерогенности,

т. к. в разных ИС СНВМ кристаллические модификации могут указываться в разных терминах (сингонии или кристаллической системы). Таким образом, имеем семь кристаллических систем: триклинная, моноклинная, ромбическая, тетрагональная, тригональная, гексагональная и кубическая.

Предлагается следующий формат данных для преставления кристаллических модификаций (с учетом данных по количественному составу) веществ. На примере кубической модификации In_2S_3 XML-документ будет следующей структуры:

```
<ModificationXML>
    <ChemicalModification>
        <Item Element = "In" value = "2" />
        <Item Element = "S" value = "3" />
        <Modification>cubic</Modification>
    </ChemicalModification>
</ModificationXML>
```

Таким образом, предложенный формат данных учитывает качественный и количественный состав веществ, а также тип кристаллической модификации.

Формат данных для представления значений свойств

При разработке форматов представления данных особое внимание уделялось способу представления значений свойств, так как было необходимо обеспечить возможность предоставления различных типов свойств (текстовые описания, графические или табличные представления). Для обеспечения кроссплатформенной возможности работы со значениями из столь широкого диапазона типов данных было принято решение использовать XML-документы для представления значений свойств. Например, для соединения $LiNbO_3$ растворимость в воде, заданная таблицей в БД «Кристалл», являющаяся функцией от температуры, представляется в виде XML-документа:

```
<root>
    <val dbid = "1" mime = "text/xml">
            <row p_TempK = "273" p_SuspName = "H&lt;sub&gt;2&lt;/sub&gt;O"
        value = "0.34" />
            <row p_TempK = "298" p_SuspName = "H&lt;sub&gt;2&lt;/sub&gt;O"
        value = "0.41" />
            <row p_TempK = "323" p_SuspName = "H&lt;sub&gt;2&lt;/sub&gt;O"
        value = "0.64" />
            <row p_TempK = "348" p_SuspName = "H&lt;sub&gt;2&lt;/sub&gt;O"
        value = "0.89" />
            <row p_TempK = "373" p_SuspName = "H&lt;sub&gt;2&lt;/sub&gt;O"
        value = "1.09" />
    </val>
</root>
```

Значения свойств хранятся в таблице DW_PropertyValues, а составной ключ IDp, IDs, IDc, IDm указывает на описываемое свойство из таблицы DW_Propeties и химическую сущность, определяемую тройкой (IDs, IDc, IDm).

Это позволяет сохранять в ХД значения свойств для химических сущностей на уровне систем, веществ и модификаций. Так, при описании значения свойства для неорганического вещества IDc содержит идентификатор соответствующего химического соединения (IDs — указывает на соответствующую химическую систему), а IDm = 0, т. е. указывает на отсутствие информации по модификации химической сущности (использование null недопустимо, в силу вхождения поля IDm в состав первичного ключа).

При работе с двоичными данными, например, с графическими рисунками или аналитическими обзорами используется представление бинарных данных в виде строки Base64 с обязательным указанием MIME-типа ресурса (например, «image/gif» или «application/pdf»). Сами значения хранятся в таблице Values, а составной ключ IDp, IDs, IDc, IDm указывает на описываемое свойство из таблицы Propeties и химическую сущность, определяемую тройкой (IDs, IDc, IDm). Это позволяет сохранять в ХД значения свойств для химических сущностей на уровне систем, веществ и модификаций. Так, при описании значения свойства для неорганического вещества IDc содержит идентификатор соответствующего химического соединения (IDs — указывает на соответствующую химическую систему), а IDm = 0, т. е. указывает на отсутствие информации по модификации химической сущности (использование null недопустимо, в силу вхождения поля IDm в состав первичного ключа) [214].

При использовании ХД в качестве источника информации для систем прогнозирования значение свойства необходимой химической сущности может быть получено двумя способами: 1) путем наложения на XML-документ специального XSLT-преобразования [162] (XML + XSLT => требуемый формат данных); 2) путем программной обработки XML документа с использованием средств организации запросов к XML-документу на языках XPath и/или XQuery. Важной особенностью является возможность выполнения данных преобразований как на стороне ХД с использованием хранимых процедур на SQL CLR (SQL Common Language Runtime — реализация размещения и запуска управляемого кода на .Net Framework в рамках СУБД Microsoft SQL Server), так и на стороне сервера приложений (например, Microsoft IIS), что улучшает масштабируемость ИС.

Преобразование к вышеуказанным форматам проводится с помощью подпрограмм, написанных с использованием среды разработки Microsoft Visual Studio 2008.

4.5.3. Процедура загрузки

В простейшем случае загрузка данных в хранилище может осуществляться через использование ADO.Net и SQL-операторов типа:

```
INSERT INTO DW_Systems (SystemID, SystemName, SystemXML)
    SELECT SystemID, SystemName, SystemXML
    FROM IMET.dbo.Crystal.DW_Systems
    WHERE SystemID NOT IN (select SystemID from DW_Systems)
```

Однако использование таких простых приемов, конечно, возможно только в рамках одной организации, при наличии непосредственного доступа к связанному серверу БД, имеющим информацию, подготовленную для загрузки в хранилище и согласованную по значениям идентификаторов.

В подавляющем большинстве случаев, такая простая загрузка данных в хранилище не может быть осуществлена, поэтому пишутся программы для загрузки XML-документов, полученных после этапа очистки преобразования данных.

Традиционная работа с данными в ADO.NET строится по следующей схеме: 1) создается соединение Connection и открывается методом Open, 2) создается объект Command, инкапсулирующий SQL-оператор, 3) выполняется SQL-оператор, 4) соединение закрывается.

Использование метода интеграции на основе хранилищ данных (ETL) для консолидации материаловедческих данных в рамках одной организации или для объединения информационных ресурсов с общей политикой доступа является общепринятой практикой. Это позволяет не только сократить временные затраты на извлечение данных из единого ХД по сравнению с разрозненными информационными источниками, но и открывает богатые возможности многомерного анализа данных. Применительно к компьютерному конструированию неорганических соединений можно рассматривать интегрированное хранилище как информационную основу для поиска новых материалов, обладающих заданными свойствами.

Краткие выводы

В главе получены следующие результаты:

- Разработаны диаграммы потоков данных (DFD) и функциональные модели методов интеграции данных (IDEF0).
- На основе полученной ER-модели разработана реляционная структура хранилища данных по свойствам неорганических веществ.
- Описаны форматы данных для представления химических сущностей и их свойств в разработанном хранилище данных.
- Разработаны и реализованы алгоритмы для извлечения, преобразования и загрузки информации по свойствам неорганических веществ в хранилище данных.

Глава 5
Использование виртуальной интеграции данных при прогнозировании свойств неорганических веществ

5.1. Подходы к интеграции информации средствами EII

При интеграции информационных систем в различных предметных областях главной задачей становится задача стандартизации. Все концепции предметной области необходимо привести к единому виду, создав некую общую схему работы. Все части интегрированной системы должны быть стандартизованы, и между всеми частями должно быть задано соответствие, и существовать отображение между ними.

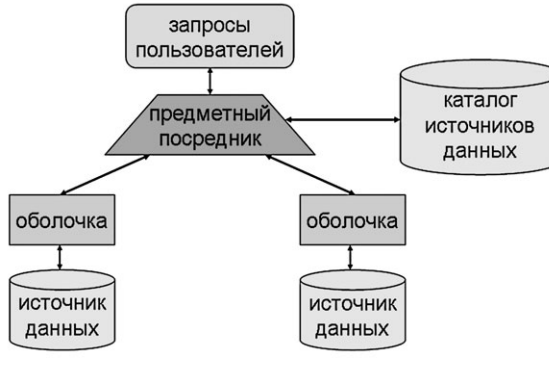


Рис. 5.1.1. Схема интеграции источников данных на основе схемы с участием предметного посредника (медиаторная схема)

В настоящее время есть ряд подходов к интеграции информации в рамках EII и множество их модификаций. Все подходы основываются на схемах с участием так называемого предметного посредника или медиатора (рис. 5.1.1).

Предметный посредник отвечает за предоставление пользователям некоторого унифицированного представления предметной области, для которой он создан. Так, все запросы пользователей поступают в предметный посредник, который отвечает за их обработку и предоставление результатов пользователям. Предметный посредник, как правило, имеет каталог источников данных, в котором содержатся сведения об интегрируемых источниках, к которым он может обращаться для получения ответов на запросы. Предметный посредник осуществляет доступ к информации в интегрируемых источниках, как правило, через специальные программные оболочки, служащие для программного согласования и называемые программными адаптерами.

Следует выделить два основных подхода к интеграции источников данных, основанных на архитектуре предметных посредников или медиаторов:

- Global-as-View (GAV) — данный подход описывает глобальную схему предметной области в терминах представлений (views) локальных схем источников данных;
- Local-as-View (LAV) — этот подход рассматривает схемы локальных источников данных как материализованные представления (materialized views) в терминах общей глобальной схемы предметной области.

Следует отметить, что существует также множество гибридных подходов, сочетающих в себе GAV и LAV (например, GLAV, BAV и др.).

Рассмотрим более подробно два основных подхода к интеграции информационных источников с использованием EII. Начнем с Global-as-View. Как уже отмечалось, данный подход описывает глобальную схему в терминах представлений (views) локальных источников. На рис. 5.1.2 глобальная схема X определена как отображение источников Y и Z. Следовательно, структура локальных источников уже «жестко» заложена в предметный посредник и активно используется им для построения ответов на запросы.

Определения отображений используются для переформулирования запросов над глобальной схемой в последовательность запросов к локальным отображениям, заданным с помощью локальных схем. Этот подход использует относительно простые методы переопределения (переформулирования) запросов, при которых ответ на запрос к общей схеме означает его развертывание (query unfolding) и переадресацию к конкретным источникам. Примерами этого подхода являются COIN [48], MOMIS[49] и IBIS [51].

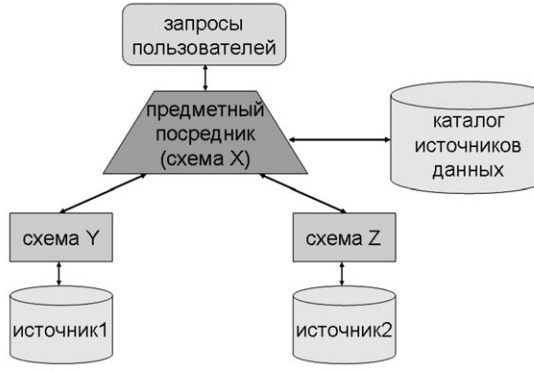


Рис. 5.1.2. Интеграция на основе принципа Global-as-View

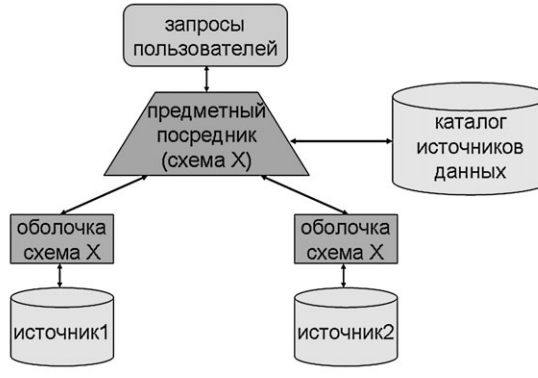


Рис. 5.1.3. Интеграция на основе принципа Local-as-View

При подходе Local-as-View [52, 53, 54] все конструкции локальной схемы определены как представления глобальной схемы. Каждый источник данных описывается одним или несколькими представлениями согласно общей схеме предметного посредника (рис. 5.1.3). Для этих преобразований поверх каждого источника создается специальный программный адаптер (wrapper) — специальная оболочка, учитывающая внутреннее представление данных в источнике и предоставляющая предметному посреднику унифицированное представление источника согласно общей схеме.

При этом подходе представления могут быть неполными, то есть они могут и не содержать записей (кортежей, если придерживаться терминов реляционных БД), удовлетворяющих определению представления, что происходит

в том случае, если источник не содержит информации, соответствующей данному представлению. Такой случай встречается довольно часто, т. к. в интегрируемых источниках данных может содержаться неполная информация, предусмотренная рамками общей глобальной схемы.

В этом случае обработка запросов над глобальной схемой сводится к перезаписи запросов с использованием отображений. Целью является переформулировка пользовательских запросов (в терминах общей схемы) в запрос, ссылающийся непосредственно на отображения всех интегрируемых источников данных, и последующий поиск ответа на него. Этот подход был, например, применен в системах Information Manifold, Agora [55].

Следует отметить, что основная проблема интеграции источников данных заключается в неразрешимости интеграции на основе отображений в общем и недостижимости интеграции на основе отображений в большинстве случаев [56]. Это связано с тем, что в источниках информации данные могут оказаться либо неполными, либо противоречивыми. В настоящее время ведется ряд работ, направленных на рассмотрение частных случаев, при которых все же можно найти ответы на запрос при наличии неполных данных. Однако следует отметить, что эти ответы могут быть разными в зависимости от гипотез, примененных к информационным источникам [57, 58, 59, 60]. Преимущество же указанных выше подходов интеграции заключается в том, что они обеспечивают общее решение, если, конечно же, оно вообще существует.

Рассмотрим кратко плюсы и минусы подходов LAV и GAV (рис. 5.1.4).

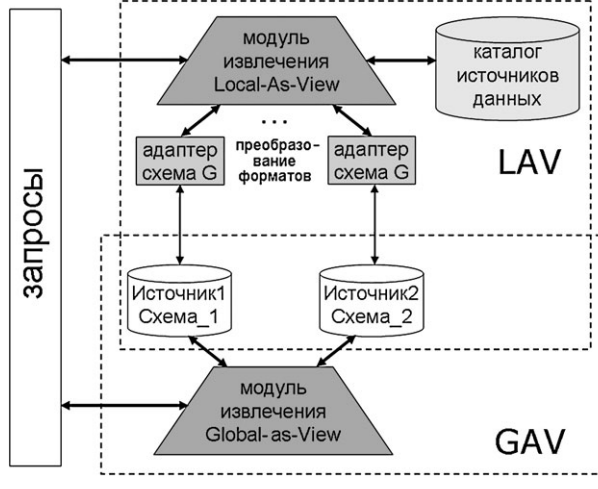


Рис. 5.1.4. Сравнение Local-as-View и Global-as-View

LAV

+ LAV-системы могут быть легко расширены за счет подключения новых источников данных, удовлетворяющих общей схеме (масштабируемость).
– Обработка запросов в LAV сложнее, чем в GAV и требует рассуждений.

GAV

– Любое добавление или изменение источника ведет к изменению всей схемы работы.
+ С другой стороны, обработка запросов более проста, чем в LAV (простота реализации).

Таким образом, построение интегрированной ИС на основе принципа LAV является более трудоемким, так как требует выработки общей глобальной схемы, в терминах которой могут быть описаны все источники данных в указанной предметной области. Разработка программных адаптеров для подключения источников данных к интегрированной ИС по принципу LAV значительно сложнее, так как каждый адаптер в данном случае выполняет полное преобразование внутренней схемы источника информации к общей глобальной схеме. В то же время, для GAV разработка таких адаптеров, строго говоря, не требуется вовсе, поскольку предметный посредник учитывает особенности информационной схемы интегрируемых источников и может осуществлять доступ непосредственно к информационному источнику.

Несмотря на трудоемкость построения интегрированной ИС по принципу LAV, у таких систем есть неоспоримое преимущество перед GAV — масштабируемость. То есть для добавления нового источника данных в интегрированную систему по принципу LAV необходимо лишь написать программный адаптер для преобразования внутренних информационных структур источника к общей информационной схеме и добавить описание данного источника в каталог источников данных интегрируемой системы. При этом подключение нового источника данных может быть выполнено на лету, без приостановки работы интегрированной ИС и изменения схем функционирования интегрированной ИС. Аналогично, в случае изменения внутренней структуры данных какого-либо источника потребуется лишь соответствующая переработка программного адаптера этого источника, в то время как интегрированная ИС не претерпит никаких изменений. Попытка же добавления нового источника данных в ИС, построенную по принципу GAV, потребует пересмотра общей схемы интегрированной ИС и, следовательно, коренных модификаций в самом предметном посреднике.

Таким образом, учитывая необходимость построения легко масштабируемой интегрированной ИС и масштабируемость интегрированных ИС, построенных по принципу LAV, принято решение использовать принцип Local-as-View для построения интегрированной ИС.

5.2. Реализация интеграции гетерогенных источников данных информационных систем

Принципы построения ИС, объединяющей гетерогенные источники данных интегрируемых ИС СНВМ, были рассмотрены в главе 3. В настоящей главе будет рассмотрена реализация интеграции источников данных ИС в рамках предложенного подхода.

Исходя из анализа доминирующих в настоящий момент технологических платформ для построения ИС, проведенного в главе 3.5 настоящей работы, было принято решение использовать технологическую платформу компании Microsoft (Windows Server + IIS 7), что позволит минимизировать затраты на разработку интегрированной ИС.

5.2.1. Описание структуры метабазы

При построении интегрированной ИС, объединяющей источники информации ИС по свойствам неорганических веществ, необходимо обеспечить хранение некоторой служебной информации. Эти данные необходимы для успешного разрешения конфликтов гетерогенности и интеграции источников данных согласно методике, описанной в разделе 3.4. Для хранения данных было принято решение использовать реляционную БД, именуемую в дальнейшем метабазой. Учитывая необходимость обработки XML-документов в рамках метабазы, для управления информационной БД, лежащей в основе интегрированной ИС, было принято решение использовать Microsoft SQL Server 2008, обладающей богатыми возможностями (.Net CLR, XPath, XQuery) по работе с XML-документами внутри БД.

В связи с тем, что интегрируемые ИС СНВМ могут пересекаться по набору свойств, а качество информации (достоверность и полнота) в каждой ИС отличается для разных свойств, целесообразно включение в структуру метабазы экспертной оценки информационных ресурсов. Экспертиза проводится высококвалифицированными специалистами, которые выставляют оценки, характеризующие качество данных в разных интегрируемых информационных системах. Это сделано для того, чтобы при

наличии информации по какому-либо физико-химическому свойству определенного вещества или химической системы в нескольких интегрируемых базах данных пользователь имел возможность выбрать наиболее достоверные и полные данные. При этом остается возможность просмотра и всей информации из разных ИС СНВМ.

При создании интегрированной ИС ключевой проблемой является разработка структуры метабазы, которая содержит ссылки на информационные ресурсы, интегрируемые по технологии LAV. Рассмотрим разработанную структуру метабазы для интеграции информационных источников по свойствам веществ, указывая, какую функциональность, опирающуюся на структуру этих таблиц, будет реализовывать интегрированная ИС (рис. 5.2.1).

Таблица Meta_DBInfo

Является главной таблицей, в которой хранится список интегрируемых источников данных ИС по свойствам неорганических веществ. Каждому подключаемому источнику данных ИС присваивается уникальный целочисленный идентификатор DBID (тип int), который является первичным ключом таблицы Meta_DBInfo. В этой таблице для каждой интегрируемой ИС содержатся учетные данные для работы с сервисами интегрированной ИС (поля Login (тип varchar(32)) и Password (тип varchar(32))). Кроме того, каждый источник данных должен предоставлять данные предметному посреднику через собственный Web-сервис. Адрес Web-сервиса и учетные данные, которые предметный посредник должен использовать для доступа к нему, хранятся в полях DBWebServiceURL (тип varchar(256)), DBWebServiceLogin (тип varchar(32)) и DBWebServicePassword (тип varchar(32)). Отметим, что новые источники данных можно подключать во время работы интегрированной ИС, добавляя соответствующие записи в эту таблицу. Также можно приостанавливать обращения предметного посредника к уже известным источникам данных, устанавливая поле Enabled (тип bit) в 0 (False).

Таблица Meta_PropertyInfo

В ней содержится список свойств, информация о которых хранится в интегрируемых источниках данных ИС. Эта таблица используется для разрешения семантических конфликтов в интегрированной ИС. Каждому свойству присваивается уникальный целочисленный идентификатор PropertyID (тип int), являющийся первичным ключом таблицы. Все Web-сервисы, предоставляющие данные из своих информационных источников, должны для обозначения свойств использовать только идентификаторы PropertyID, содержащиеся в этой таблице. Поле Name (тип varchar(256)) содержит название свойства, а поля SynonymsXML и SynonymsString используются для хранения списка синонимов к названию свойства в поле Name и задействованы при разрешении семантических конфликтов. Важным является

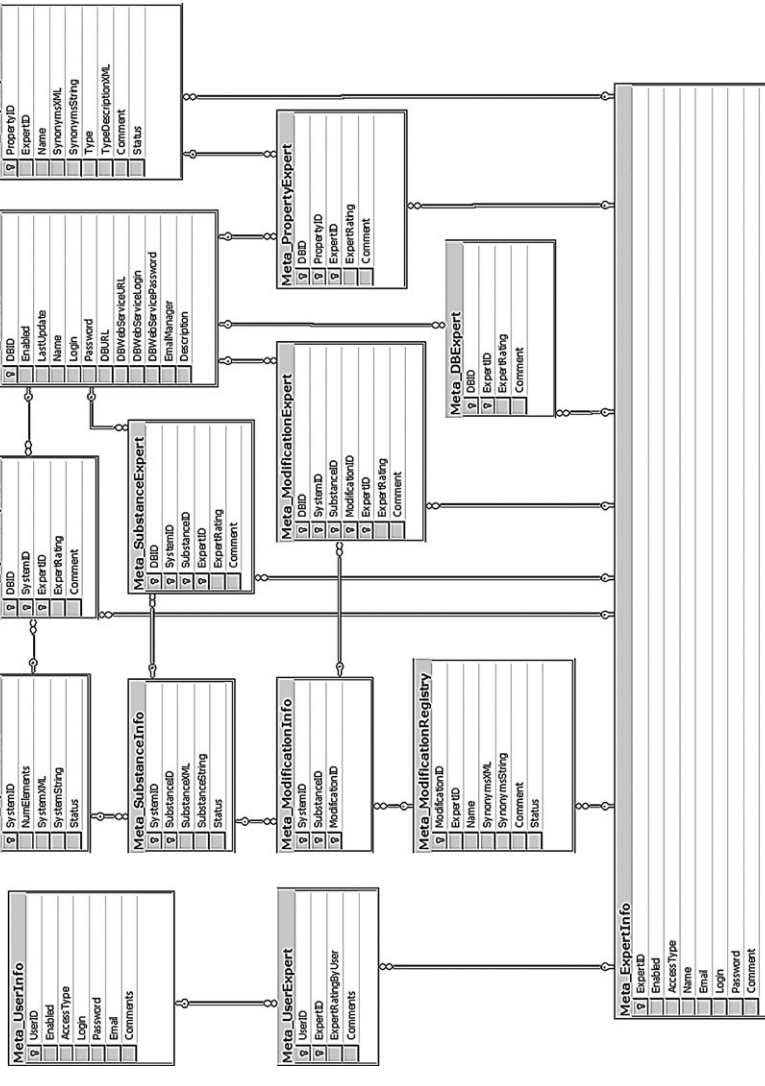


Рис. 5.2.1. Структура метабазы информационной системы, интегрирующей источники данных информационных систем по свойствам неорганических веществ

поле Status (тип int), которое используется для хранения статуса глобального идентификатора, который обсуждался в разделе 3.4. При этом статусу «надежный» соответствует 1, а статусу «ненадежный» соответствует 0. Поле Status присутствует и в других таблицах, а его назначение всегда связано с надежностью глобальных идентификаторов, присутствующих в соответствующих таблицах.

Таблица Meta_SystemInfo

Предназначена для хранения списка химических систем (наборов обозначений химических элементов, образующих химическую систему), зарегистрированных в метабазе. Порядок следования химических элементов, естественно, значения не имеет. Каждой химической системе соответствует уникальный идентификатор SystemID (тип int), являющийся первичным ключом таблицы и использующийся при ссылках на соответствующую химическую сущность. Эта таблица используется для сквозной нумерации всех химических систем в интегрированной ИС. Поле SystemXML (тип xml) содержит описание химической системы в оговоренном XML-schema [165] формате — это фактически множество химических элементов (рис. 5.2.2). Можно сказать, что таблица Meta_SystemInfo задает множество S, описанное в главе 3.4, т. е. каждая строка таблицы есть элемент множества: $s = \{e_1, e_2, .., e_n\}$.

```
<?xml version="1.0" encoding="windows-1251" ?>
- <ChemicalSystem>
    <Element>As</Element>
    <Element>Ga</Element>
  </ChemicalSystem>
```

Рис. 5.2.2. Пример XML-документа, описывающего химическую систему Ga–As (снимок экрана из Microsoft IE)

Таблица Meta_SubstanceInfo

Предназначена для хранения списка химических веществ, зарегистрированных в метабазе. Согласно принятой в главе 3.4 модели, вещество c может быть представлено кортежем $< s, f >$, т. е. дополнительно к набору химических элементов определено и количественное вхождение каждого элемента в состав вещества. Каждому химическому веществу ставится в соответствие пара уникальных целочисленных идентификаторов (SystemID, SubstanceID), являющихся ключом таблицы. Поле SubstanceXML (тип xml) содержит описание химического вещества в оговоренном XML-schema [165] формате (рис. 5.2.3).

```
<?xml version="1.0" encoding="windows-1251" ?>
- <ChemicalSubstanceComposition>
    <Item Element="In" value="2" />
    <Item Element="S" value="3" />
  </ChemicalSubstanceComposition>
```

Рис. 5.2.3. Пример XML-документа, описывающего химическое вещество In_2S_3 (снимок экрана из Microsoft IE)

Таблица Meta_ModificationRegistry

Предназначена для хранения возможных обозначений кристаллических модификаций. Основное ее назначение — разрешение семантических конфликтов в обозначениях модификаций, используемых в различных интегрируемых ИС. В этой таблице возможным обозначениям кристаллических модификаций ставится в соответствие целочисленный идентификатор ModificationID (тип int), являющийся первичным ключом таблицы. Поле Name (тип varchar(256)) содержит название модификации, а поля SynonymsXML и SynonymsString используются для хранения списка синонимов к названию модификации в поле Name и задействованы при разрешении семантических конфликтов. Поле Status (тип int) используется для хранения статуса глобального идентификатора, который обсуждался в разделе 3.4. При этом статусу «надежный» соответствует 1, а статусу «ненадежный» соответствует 0. Поле SynonymsXML (тип xml) содержит список синонимов обозначения кристаллической модификации вещества в оговоренном XML-формате (рис. 5.2.4).

```
<?xml version="1.0" encoding="windows-1251" ?>
- <SynonymsList>
    <Synonym>Ромбоэдрическая</Synonym>
    <Synonym>рэ</Synonym>
    <Synonym>Rhombohedral</Synonym>
  </SynonymsList>
```

Рис. 5.2.4. Пример XML-документа, описывающего список синонимов ромбоэдрической модификации (снимок экрана из Microsoft IE)

Таблица Meta_ModificationInfo

Предназначена для хранения списка кристаллических модификаций химических веществ, зарегистрированных в метабазе. Ключом таблицы является тройка идентификаторов SystemID, SubstanceID, ModificationID.

Фактически, эта тройка идентификаторов соответствует тройке (s,c,m), где $s \in S, c \in C, m \in M$, рассмотренной в разделе 3.4.

Таблица Meta_ExpertInfo

В ней хранится список экспертов, которые имеют право осуществлять оценку качества интегрируемых данных. Ключом таблицы является целочисленный идентификатор эксперта ExpertID (тип int). Поле Enabled (тип bit) указывает, активен ли эксперт, т. е. может ли он осуществлять вход в ИС и оценку качества данных (1) или временно доступ для него закрыт (0). Учетные данные экспертов хранятся в полях Login (тип varchar(32)) и Password (тип varchar(32)). Поле AccessType (тип int) содержит спецификацию прав доступа эксперта к системе и может принимать следующие значения:

- 0 — эксперт имеет только право оценивать качество данных. Фактически под этим подразумевается внесение изменений (только от своего имени) в таблицы Meta_DBExpert, Meta_PropertyExpert, Meta_SystemExpert, Meta_SubstanceExpert, Meta_ModificationExpert.
- 1 — эксперт имеет право оценивать качество данных (как при AccessType = 0) и корректировать работу подсистемы разрешения семантических конфликтов. Под этим понимается управление полями Status таблиц Meta_PropertyInfo, Meta_SystemInfo, Meta_SubstanceInfo, Meta_ModificationRegistry и внесение в них соответствующих изменений. Также допускается исправление списка синонимов для соответствующих сущностей в этих таблицах.

Таблица Meta_SystemExpert

В ней содержатся экспертные оценки качества данных для химических систем, описанных в источниках данных интегрируемых ИС. Ключом таблицы является связка полей DBID, SystemID, ExpertID. Это означает, что экспертом ExpertID оценивается качество данных для химической системы с идентификатором SystemID в интегрируемой системе DBID. Оценка содержится в поле ExpertRating (тип float) и может быть в интервале [0; 10]. Чем больше значение поля, тем выше, по мнению эксперта, качество данных. Если экспертная оценка экспертом ExpertID для химической системы SystemID интегрируемой ИС DBID не задана, то она считается равной 1.

Таблица Meta_SubstanceExpert

Хранит экспертные оценки качества данных для химических веществ, описанных в источниках данных интегрируемых ИС. Ключом таблицы является связка полей DBID, SystemID, SubstanceID, ExpertID. Назначение полей аналогично таблице Meta_SystemExpert.

Таблица Meta_ModificationExpert

Хранит экспертные оценки качества данных для модификаций химических веществ, описанных в источниках данных интегрируемых ИС. Ключом таблицы является связка полей DBID, SystemID, SubstanceID, ModificationID, ExpertID. Назначение полей аналогично таблице Meta_SystemExpert.

Таблица Meta_DBExpert

Хранит экспертные оценки качества данных в целом в интегрируемом источнике данных ИС. Ключом таблицы является связка полей DBID, ExpertID. Назначение полей аналогично таблице Meta_ SystemExpert.

Таблица Meta_PropertyExpert

В ней хранятся экспертные оценки качества данных для свойств, описываемых в интегрируемом источнике данных ИС. Ключом таблицы является связка полей DBID, PropertyID, ExpertID. Назначение полей аналогично таблице Meta_SystemExpert.

Таким образом, таблицы Meta_DBExpert, Meta_PropertyExpert, Meta_ SystemExpert, Meta_SubstanceExpert, Meta_ModificationExpert содержат экспертные оценки качества информации в интегрируемых источниках данных ИС. Таблицы приведены в порядке возрастания приоритетов экспертных оценок. Например, если рассматривается ширина запрещенной зоны вещества In_2S_3, и определены экспертные оценки для интегрируемой ИС (таблица Meta_DBExpert) рассматриваемого свойства (таблица Meta_PropertyExpert) и для химической системы SystemID (таблица Meta_SystemExpert), соответствующей системе In–S, то будет использоваться экспертная оценка для химической системы SystemID из таблицы Meta_SystemExpert.

Таблица Meta_UserInfo

Предназначена для хранения списка зарегистрированных пользователей интегрированной ИС. Каждому пользователю присваивается уникальный целочисленный идентификатор UserID (тип int), являющийся первичным ключом таблицы. Поле Enabled (тип bit) указывает, активен ли пользователь, т. е. может ли он осуществлять вход в ИС, т. е. ее использование (1), или временно доступ для него закрыт (0). Учетные данные пользователей хранятся в полях Login (тип varchar(32)) и Password (тип varchar(32)). Поле AccessType (тип int) в настоящее время зарезервировано и не используется, т. к. все активные пользователи обладают одинаковыми правами — возможностью просмотра всей информации, содержащейся в интегрируемых ИС.

Таблица Meta_UserExpert

Предназначена для хранения пользовательского рейтинга экспертов. Каждый пользователь может выставлять свой рейтинг экспертам, оценивающим качество данных в интегрированных ИС. Тем самым пользователь может повысить или понизить значимость оценок конкретных экспертов при выдаче данных интегрированной ИС. По умолчанию оценки всех экспертов обладают одинаковым приоритетом, т. е. эксперты вносят одинаковый вклад в оценку данных. Это делается для того, чтобы пользователь получал наиболее достоверные с точки зрения экспертов данные. Так как данные, отвечающие критериям пользовательского запроса, могут содержаться сразу в нескольких интегрируемых ИС, пользователю выдаются все данные в порядке уменьшения обобщенных экспертных оценок. То есть интегрированная ИС выводит сначала наиболее достоверные данные, а затем наименее достоверные по мнению экспертов данные. Таким образом, в зависимости от пользовательских оценок, выставляемых разным экспертам, интегрированная ИС может выводить данные в разном порядке, зависящем от пользовательского рейтинга экспертов, который у каждого пользователя свой.

Первичным ключом таблицы Meta_UserExpert является связка полей UserID, ExpertID означающих, что пользователем с идентификатором UserID задается рейтинг эксперту, определяемому идентификатором ExpertID. Сам рейтинг представляется вещественным числом в интервале [0; 10] и хранится в поле ExpertRatingByUser (тип float). Если для эксперта не задан рейтинг, то он считается равным 1. То есть если пользователь не выставил рейтинги экспертам, то все рейтинги по умолчанию будут равны 1, следовательно, все эксперты будут вносить одинаковый вклад в оценку качества данных.

Таким образом, рассмотренные таблицы по их назначению можно условно отнести к нескольким группам:

- **Meta_DBInfo** — корневая таблица, содержащая информацию об интегрируемых ИС;
- **Meta_ExpertInfo, Meta_UserInfo, Meta_UserExpert** — таблицы, содержащие информацию об экспертах и пользователях ИС, их правах и пользовательских рейтингах экспертов;
- **Meta_SystemInfo, Meta_SubstanceInfo, Meta_ModificationInfo, Meta_ModificationRegisrty, Meta_PropertyInfo —** таблицы, предназначенные для разрешения семантических конфликтов на уровне: химических систем, веществ, модификаций и свойств;
- **Meta_DBExpert, Meta_PropertyExpert, Meta_SystemExpert, Meta_SubstanceExpert, Meta_ModificationExpert** — таблицы, содержащие экспертные оценки качества информации в интегрируемых источниках данных ИС.

5.2.2. Расчет достоверности информации, основанный на экспертных оценках

Предположим, что вся информация в интегрируемых источниках данных оценивается экспертами, входящими в таблицу Meta_ExpertInfo. Таким образом, имеем множество экспертов $E = \{e_1, e_2, ..., e_n\}$, где e_i — эксперт, содержащийся в таблице Meta_ExpertInfo. При этом $|E| = n$, т. е. n — количество экспертов, оценивающих качество данных.

Пусть есть некоторая информация, оцениваемая экспертами, например, это может быть качество данных о химическом веществе In_2S_3, информация о котором хранится в ИС «Bandgap». Каждый эксперт оценивает качество данных, выставляя свою оценку x_i. Как отмечалось выше, экспертные оценки качества данных могут содержаться в таблицах Meta_DBExpert, Meta_PropertyExpert, Meta_SystemExpert, Meta_SubstanceExpert, Meta_ModificationExpert. Таблицы приведены в порядке возрастания приоритетов экспертных оценок. Напомним, что оценка x_i — это любое число из интервала [0; 10], если оценка не выставлена, то она считается равной 1 (т. е. $x_i = 1$). Так получается множество экспертных оценок для оцениваемой информации $X = \{x_1, x_2, ..., x_n\}$.

Обозначим через U множество пользователей системы, т. е. $U = \{u_1, u_2, ..., u_m\}$, где u_i — пользователь, содержащийся в таблице Meta_UserInfo. Как отмечалось выше, каждый пользователь может в большей или меньшей степени доверять экспертам. Чтобы учесть эти предпочтения, каждый пользователь выставляет свои оценки экспертам. Т. к. $|E| = n$, то пользователь может выставить n оценок, сформировав тем самым множество $A = \{a_1, a_2, ..., a_n\}$, где a_i — рейтинг, выставленный пользователем i-му эксперту, a_i может принимать значения в интервале [0; 10]. Если пользователь явно не задает рейтинг эксперта, то он считается равным 1 (т. е. $a_i = 1$).

Интегрированная ИС при ответе на запрос пользователя выводит данные, упорядоченные по уменьшению степени их достоверности. Степень достоверности информации D, приводимой источниками данных, рассчитывается при наличии экспертов ($n > 0$) для каждого пользователя по формуле:

$$D = \frac{1}{n}(a_1 \cdot x_1 + a_2 \cdot x_2 + ... + a_n \cdot x_n) = \frac{1}{n}\sum_{i=1}^{n} a_i \cdot x_i .$$

Здесь n — количество экспертов, оценивающих качество данных, a_i — рейтинг, выставленный пользователем i-му эксперту, x_i — оценка качества данных, выставленная i-м экспертом.

Следует заметить, что если эксперты, оценивающие качество данных не введены, другими словами, если множество E пусто, то вышеприведенной формулой пользоваться нельзя (т. к. при $n = 0$ произойдет деление на ноль), а достоверность информации D принимается равной единице, т. е. $D = 1$.

Отметим, что пользователь может не учитывать оценки, выставляемые определенными экспертами. Для этого он выставляет соответствующим экспертам рейтинг $a_i = 0$. Если же он хочет повысить значимость оценок определенного эксперта, то выставляется рейтинг $a_i > 1$.

Несложно видеть, что если эксперты не оценивают качество данных, а пользователи не выставляют экспертам рейтинг, то $D \equiv 1$, т. к. $a_i = 1$ и $x_i = 1$. В таком случае все данные из интегрируемых источников будут считаться одинаково достоверными, и, следовательно, предметный посредник будет выводить их в произвольном порядке. Таким образом, интегрированная ИС может успешно функционировать и без экспертной оценки качества интегрируемых данных. В этом случае все данные из интегрируемых источников считаются одинаково достоверными.

5.2.3. Разработка программных адаптеров интегрируемых информационных систем

При интеграции данных ИС предметный посредник (медиатор) осуществляет доступ к информации в интегрируемых источниках через специальные программные оболочки, которые служат для программного согласования форматов данных предметного посредника с интегрируемыми источниками и называются программными адаптерами (см. рис. 1.4).

Для реализации интегрированной ИС был выбран подход Local-as-View, описывающий все конструкции локальной схемы данных информационного источника через представления глобальной схемы. Для осуществления кроссплатформенного взаимодействия было принято решение использовать технологию Web-сервисов. Таким образом, все программные адаптеры интегрируемых информационных систем реализуются в качестве Web-сервисов, доступных по протоколу SOAP. Это позволяет успешно разрешать платформенные конфликты, т. к. Web-сервисы могут быть разработаны на любой современной программной платформе, на которой реализованы интегрируемые источники данных.

Учитывая то, что был выбран подход Local-as-View, программные адаптеры предоставляют во внешнюю среду данные в едином формате, оговоренном при описании общей схемы предметной области. При этом адаптеры должны для обозначения сущностей предметной области использовать только глобальные идентификаторы, которые были получены

от предметного посредника (см. раздел 5.2.4). Таким образом, все Web-сервисы должны иметь одинаковое, стандартизированное в рамках общей схемы WSDL-описание, что обеспечит унифицированную работу со всеми такими Web-сервисными адаптерами со стороны предметного посредника.

Учитывая общую схему предметной области, было разработано общее описание Web-сервисов программных адаптеров на языке WSDL [178], которому должны удовлетворять все Web-сервисы адаптеров интегрируемых ИС.

Рассмотрим кратко назначение основных методов Web-сервиса, играющего роль программного адаптера интегрируемой ИС.

Методы **GetMataInfo** и **Supports** используются для получения версии программного адаптера и набора поддерживаемых функций соответственно. Эти методы предназначены для реализации механизма поддержки версий программных адаптеров и будут использоваться по мере развития интегрированной информационной системы для обеспечения корректной работы предметного посредника с различными версиями адаптеров информационных систем. В настоящее время предметным посредником версии 1.0 поддерживается только версия 1.0 программных адаптеров.

Метод **GetData** используются для безопасного извлечения данных из интегрируемого информационного источника согласно переданным параметрам запроса. Метод осуществляет передачу запрашиваемых данных только в том случае, если учетная запись, предъявленная при его вызове, является учетной записью пользователя соответствующего информационного ресурса. Этот метод поддерживает всю функциональность, связанную с извлечением данных из источника, преобразованием их к формату, определенному общей схемой, включая использование глобальных идентификаторов сущностей предметной области и выдачу их предметному посреднику. В версии 1.0 адаптера метод GetData поддерживает три подкоманды: Get_AllList, Get_PropertiesList и Get_PropertiesValues. Рассмотрим кратко назначение приведенных подкоманд.

Подкоманда **Get_AllList** метода GetData предназначена для запроса химических сущностей, освещаемых в рамках интегрируемого источника. Предусмотрен механизм фильтрации возвращаемых данных по химической системе, веществу и модификации. Критерий фильтрации задается в качестве входного аргумента метода GetData и является XML-документом оговоренного формата. Пример XML-документа, возвращающего предметному посреднику отфильтрованный список содержимого информационного ресурса, приведен на рис. 5.2.5. В документе описано химическое вещество с глобальным идентификатором (SystemID = 1, SubstanceID = 86, ModificationID = 0) и две его модификации с идентификаторами (SystemID = 1, SubstanceID = 86, ModificationID = 5) и (SystemID = 1, SubstanceID = 86, ModificationID = 7) соответственно. При этом вещество

описано согласно шаблону для записи химических веществ $(s, c, null)$, а модификации вещества описаны согласно шаблону для записи химических модификаций (s, c, m) (см. раздел 3.4.2).

```
<?xml version="1.0" encoding="windows-1251" ?>
- <root>
    <row SystemID="1" SubstanceID="86" ModificationID="0" />
    <row SystemID="1" SubstanceID="86" ModificationID="5" />
    <row SystemID="1" SubstanceID="86" ModificationID="7" />
  </root>
```

Рис. 5.2.5. Пример XML-документа, описывающего содержимое информационного источника (снимок экрана из Microsoft IE)

Подкоманда **Get_PropertiesList** метода GetData предназначена для запроса набора свойств, освещаемого в рамках интегрируемого источника с возможностью фильтрации, задаваемой аналогично фильтрации в подкоманде Get_AllList в качестве входного аргумента метода GetData. Пример XML-документа, возвращающего предметному посреднику отфильтрованный список свойств, рассмотренных в информационном ресурсе, приведен на рис. 5.2.6. Как видно, в документе описан набор свойств с глобальными идентификаторами PropertyID, полученными от предметного посредника.

```
<?xml version="1.0" encoding="windows-1251" ?>
- <root>
    <row PropertyID="5" Name="Удельная теплоемкость" />
    <row PropertyID="6" Name="Плотность" />
    <row PropertyID="7" Name="Твердость" />
    <row PropertyID="8" Name="Растворимость" />
    <row PropertyID="9" Name="Температура плавления" />
    <row PropertyID="10" Name="Температура Кюри" />
  </root>
```

Рис. 5.2.6. Пример XML-документа, описывающего свойства, освещенные в информационном источнике (снимок экрана из Microsoft IE)

Полный список свойств, освещенных в ИС «Кристалл», приведен в табл. 5.1.

Подкоманда **Get_PropertiesValues** метода GetData предназначена для запроса значений свойств, освещаемых в рамках интегрируемого источника

с возможностью фильтрации, задаваемой аналогично фильтрации в подкоманде Get_AllList с помощью XML-документа. Пример XML-документа, возвращающего предметному посреднику отфильтрованный список значений свойств в информационном ресурсе, приведен на рис. 5.2.7. Как видно, в документе описаны значения акустооптических свойств (PropertyID = 27) из БД «Кристалл» (DBPropID = ”crystal.*”) для химического вещества GaAs (SystemID = 1, SubstanceID = 86, ModificationID = 0).

Таблица 5.1. Свойства веществ, описанные в ИС «Кристалл»

№	Свойство
1	Теплоемкость
3	Аналитический обзор
4	Состав соединения
5	Удельная теплоемкость
6	Плотность
7	Твердость
8	Растворимость
9	Температура плавления
10	Температура Кюри
11	Характеристика кристаллической структуры
12	Параметры элементарной ячейки
13	Тепловое расширение
14	Теплопроводность
15	Диэлектрическая проницаемость
16	Тангенс угла диэлектрических потерь
17	Пьезоэлектрические коэффициенты
18	Коэффициенты электромеханической связи
19	Упругие постоянные
20	Полоса пропускания
21	Показатели преломления
22	Коэффициенты Селмейера
23	Коэффициенты линейного электрооптического эффекта
24	Нелинейные оптические свойства
25	Пьезооптические и упругооптические коэффициенты
26	Распространение и затухание упругих волн
27	Акустооптические свойства
28	Литература

```
<?xml version="1.0" encoding="windows-1251" ?>
- <root>
  - <row PropertyID="27" Name="Акустооптические свойства" DBPropID="crystal.AcOpTabl">
    - <PropertyXML>
        <row SystemID="1" SubstanceID="86" ModificationID="0" WaveLeng="1.15" Nzv="[100]" Uzv="[010]" E="произв."
          M1="155" M2="46.3" M3="49.2" Reference="<B>Ярив А.,Юх П.</B> <I>Оптические волны в кристаллах, М.:
          Мир, 1987, 616 с.</I> //" />
        <row SystemID="1" SubstanceID="86" ModificationID="0" WaveLeng="1.15" Nzv="[110]" Uzv="[110]" Nsv="[110]"
          E="[110]" M1="925" M2="104" M3="179" Reference="<B>Ярив А.,Юх П.</B> <I>Оптические волны в
          кристаллах, М.: Мир, 1987, 616 с.</I> //" />
        <row SystemID="1" SubstanceID="86" ModificationID="0" WaveLeng="1.153" Nzv="[100]" Uzv="[010]" E="произв."
          M1="155" M2="46.3" M3="49.2" Reference="<B>Блистанов А.А.,Бондаренко В.С.,Переломова
          Н.В.,Стрижевская Ф.И.,Чкалова В.В.,Шаскольская М.П.</B> <I>Акустические кристаллы. Справочник. М.:
          Наука, 1982, 632 с.</I> //" />
        <row SystemID="1" SubstanceID="86" ModificationID="0" WaveLeng="1.153" Nzv="[110]" Uzv="[110]" E="[110]"
          M1="925" M2="104" M3="179" Reference="<B>Блистанов А.А.,Бондаренко В.С.,Переломова Н.В.,Стрижевская
          Ф.И.,Чкалова В.В.,Шаскольская М.П.</B> <I>Акустические кристаллы. Справочник. М.: Наука, 1982, 632
          с.</I> //" />
      </PropertyXML>
    </row>
  </root>
```

Рис. 5.2.7. Пример XML-документа, описывающего значения свойства в информационном источнике (снимок экрана из Microsoft IE)

Следует отметить, что у разных свойств в разных информационных источниках присутствуют различные атрибуты и, следовательно, структура XML-узла PropertyXML не является жестко фиксированной. Для того, чтобы отобразить эти данные конечному пользователю или подать на вход СППР, используют XSLT-преобразования, приводящие данные к требуемому виду.

5.2.4. Разработка предметного посредника

Предметный посредник является точкой входа в интегрированную ИС и реализует ответы на запросы с использованием информации, размещенной в интегрированных источниках, доступ к которым он осуществляет через программные адаптеры. Предметный посредник реализован в виде Web-сервиса, доступного по адресу https://meta.imet-db.ru/eii/Service.asmx с использованием протокола SOAP. Реализация предметного посредника в качестве Web-сервиса, оперирующего XML-документами, позволяет успешно осуществлять доступ к нему с любой современной программной платформы, на которой возникнет необходимость в использовании данных из интегрированной ИС.

Было разработано общее описание Web-сервиса предметного посредника на языке WSDL [179]. Используя это WSDL-описание можно автоматически создать во многих современных программных средах прокси-классы для осуществления доступа к предметному посреднику и, тем самым, использовать интегрированную ИС по свойствам неорганических веществ.

Рассмотрим кратко назначение основных методов Web-сервиса предметного посредника интегрируемой ИС. Метод **GetMataInfo** используется для получения версии предметного посредника и предназначен для реализации механизма поддержки версий. Текущей версией предметного посредника является версия 1.0.

Метод **ProcessCommand** используется программными адаптерами интегрируемых ИС для определения глобальных идентификаторов сущностей предметной области. Другими словами, этот метод используется для разрешения синтаксических, структурных и семантических конфликтов, описанных в разделах 2.6.2 и 2.6.3 соответственно. Этот метод может быть вызван только с использованием учетных данных (поля Login и Password), определенных в таблице DBInfo. Таким образом, правом вызова этого метода обладают только интегрируемые ИС. В качестве входного аргумента в этот метод передается XML-документ, содержащий описание сущностей, глобальные идентификаторы которых необходимо получить программному адаптеру интегрируемого источника, чтобы в дальнейшем осуществлять взаимодействие с предметным посредником. Пример XML-документа показан на рис. 5.2.8.

```
<?xml version="1.0" encoding="windows-1251" ?>
- <MetaInfo date="2006-03-15T13:40" version="1.0">
  - <ModificationList>
      <Modification id="тр" Modification="тр" />
    </ModificationList>
  - <PropertyList>
      <Property id="6" Name="Плотность" />
    </PropertyList>
  - <ChemicalSubstanceList>
      <ChemicalSubstance id="1" fromHTML="LiNbO<sub>3</sub>" />
    </ChemicalSubstanceList>
  </MetaInfo>
```

Рис. 5.2.8. Пример XML-документа, подаваемого на вход предметного посредника для получения глобальных идентификаторов сущностей (снимок экрана из Microsoft IE)

Предметный посредник производит обработку всех узлов, соответствующих сущностям предметной области и расположенных в XML-документе согласно XPath-выражениям:

- «/MetaInfo/ChemicalSystemList/ChemicalSystem» — путь для химических систем (на основе таблицы Meta_SystemInfo);
- «/MetaInfo/ChemicalSubstanceList/ChemicalSubstance» — путь для химических веществ и их модификаций (на основе таблиц Meta_SystemInfo, Meta_SubstanceInfo и Meta_ModificationInfo);
- «/MetaInfo/ModificationList/Modification» — путь для разрешения семантических конфликтов в обозначениях кристаллических модификаций (на основе таблицы Meta_ModificationRegistry);
- «/MetaInfo/PropertyList/Property» — путь для разрешения семантических конфликтов в обозначениях свойств информационного источника (на основе таблицы Meta_PropertyInfo).

В результате обработки происходит определение глобальных идентификаторов соответствующих сущностей на основе записей в таблицах метабазы и соответствующих им статусов. После этого данная информация записывается в узлы XML-документа, соответствующие сущностям, а сам XML-документ возвращается в качестве ответа Web-сервиса. Пример такого документа-ответа, соответствующего документу-запросу с рис. 5.2.8, приведен на рис. 5.2.9.

Следует отметить, что данный XML-документ соответствует документу-запросу на рис. 5.2.8 и содержит глобальные идентификаторы заданные в атрибутах SystemID, SubstanceID, ModificationID, PropertyID и соответствующие им статусы глобальных идентификаторов в атрибутах Status_SystemID, Status_SubstanceID, Status_ModificationID, Status_PropertyID.

```
<?xml version="1.0" encoding="windows-1251" ?>
- <MetaInfo date="2006-03-15T13:40" version="1.0">
  - <ModificationList>
      <Modification id="тр" Modification="тр" ModificationID="1"
        Status_ModificationID="0" />
    </ModificationList>
  - <PropertyList>
      <Property id="6" Name="Плотность" PropertyID="6" Status_PropertyID="0" />
    </PropertyList>
  - <ChemicalSubstanceList>
      <ChemicalSubstance id="1" fromHTML="LiNbO<sub>3</sub>" SystemID="7"
        Status_SystemID="0" SubstanceID="5" Status_SubstanceID="0"
        ModificationID="0" Status_ModificationID="1" />
    </ChemicalSubstanceList>
  </MetaInfo>
```

Рис. 5.2.9. Пример XML-документа, передаваемого в качестве ответа Web-сервиса на запрос глобальных идентификаторов (снимок экрана из Microsoft IE)

Необходимо отметить, что в узлах документа-запроса, соответствующих путям «/MetaInfo/ChemicalSystemList/ChemicalSystem» и «/MetaInfo/Chemical SubstanceList/ChemicalSubstance» могут содержаться сущности, описанные различными способами. Например, химическое вещество может быть представлено в качестве HTML-формулы, заданной в атрибуте fromHTML и в качестве иерархической структуры описывающей соответствующее вещество.

Методы **childEIIService_GetMetaInfo**, **childEIIService_Supports**, **childEIIService_GetData** предметного посредника предназначены для непосредственных вызовов методов GetMetaInfo, Supports, GetData программных адаптеров источников данных интегрируемых ИС. При этом для вызова этих методов должна использоваться учетная запись пользователя интегрированной ИС (из таблицы UserInfo), а нужный источник идентифицируется соответствующим учетным именем (поле Login из таблицы DBInfo). Эти методы могут использоваться в том случае, когда пользователям предметного посредника требуется непосредственно извлечь данные из интегрируемых источников без их обработки предметным посредником. Эта возможность вводится для обеспечения гибкости в работе пользователей и реализации ими сценариев взаимодействия с интегрируемыми ИС, которые не предусмотрены предметным посредником.

Метод **GetAllCommulativeData** предметного посредника предназначен для извлечения информации, содержащейся в интегрированной ИС, объединяемой предметным посредником. При вызове этого метода предметный посредник опрашивает все информационные источники интегрируемых ИС, анализирует их ответы и создает XML-документ, содержащий результирующий ответ предметного посредника на запрос пользователя.

Для успешного вызова метода необходимо передать учетные данные пользователя интегрированной ИС, один из трех типов команды Get_AllList, Get_PropertiesList или Get_PropertiesValues, которые были рассмотрены при описании метода **GetData** программного адаптера интегрируемой ИС. Также на вход метода подается XML-документ с описанием параметров запроса.

Рассмотрим результаты работы этого метода на примере вызова команды Get_PropertiesValues с передачей XML-документа с описанием параметров запроса, показанного на рис. 5.2.10.

```
<?xml version="1.0" encoding="windows-1251" ?>
- <root>
    <ExpertRating />
    <DescribeChemicalEntities />
  - <PropertyID>
      <item value="27" />
    </PropertyID>
  - <SystemID NotIn="0">
      <item value="1" />
    </SystemID>
  - <SubstanceID NotIn="0">
      <item value="86" />
    </SubstanceID>
  - <ModificationID NotIn="1">
      <item value="1" />
      <item value="2" />
    </ModificationID>
  </root>
```

Рис. 5.2.10. Пример XML-документа, задающего параметры запроса при вызове метода GetAllCommulativeData предметного посредника (снимок экрана из Microsoft IE)

Отметим, что так как используется команда GetAllCommulativeData, то в данном случае пользователь запрашивает значения свойств химических сущностей. Рассмотрим кратко параметры запроса для команды GetAllCommulativeData, передаваемые с помощью XML-документа, показанного на рис. 5.2.10. Поскольку в интегрированной ИС может содержаться большое количество разных свойств и химических сущностей, запрос уточняется с помощью XML-документа. В этом документе присутствует узел, соответствующий XPath-пути «/root/PropertyID», отвечающий за фильтрацию запрашиваемых свойств. В нашем примере в этом узле присутствует только один дочерний узел item со значением 27, это означает, что запрашиваются только значения свойства с глобальным идентификатором PropertyID = 27, что соответствует акустооптическим свойствам.

Узлы, соответствующие XPath-путям «/root/SystemID», «/root/SubstanceID» и «/root/ModificationID» отвечают за фильтрацию запрашиваемых химических сущностей. При этом узел «/root/SystemID» фильтрует сущности на уровне химических систем, «/root/SubstanceID» — на уровне химических веществ и «/root/ModificationID» — на уровне химических модификаций. В нашем примере на рис. 5.2.10 накладывается фильтр по химической системе с глобальным идентификатором SystemID = 1 (химическая система As–Ga), по химическому веществу с глобальным идентификатором SystemID = 1, SubstanceID = 86 (химическое вещество GaAs). По химической модификации накладывается фильтр, указывающий, что глобальный идентификатор химической модификации ModificationID не должен принимать значения 1 и 2 (т. к. атрибут NotIn узла «/root/ModificationID» равен единице).

Отметим наличие еще двух узлов, соответствующих XPath-путям «/root/ExpertRating», «/root/DescribeChemicalEntities». Узел ExpertRating указывает предметному посреднику на необходимость добавления в XML-документ с ответом значений экспертного рейтинга приведенной информации.

```
<?xml version="1.0" encoding="windows-1251" ?>
- <root>
  - <row PropertyID="27" Name="Акустооптические свойства"
      DBPropID="crystal.AcOpTabl" Rating="1.0">
    - <PropertyXML>
      - <row SystemID="1" SubstanceID="86" ModificationID="0" WaveLeng="1.15"
          Nzv="[100]" Uzv="[010]" E="произв." M1="155" M2="46.3" M3="49.2"
          Reference="<B>Ярив А.,Юх П.</B> <I>Оптические волны в кристаллах,
          М.: Мир, 1987, 616 с.</I> //" Rating="1.0">
        - <SystemXML>
          - <ChemicalSystem>
              <Element>As</Element>
              <Element>Ga</Element>
            </ChemicalSystem>
          </SystemXML>
        - <SubstanceXML>
          - <ChemicalSubstanceComposition>
              <Item Element="As" value="1" />
              <Item Element="Ga" value="1" />
            </ChemicalSubstanceComposition>
          </SubstanceXML>
          <ModificationName />
        </row>
      </PropertyXML>
    </row>
</root>
```

Рис. 5.2.11. Фрагмент XML-документа, содержащего ответ предметного посредника интегрируемой ИС на запрос значений свойств (снимок экрана из Microsoft IE)

Экспертный рейтинг добавляется к соответствующим узлам результирующего документа в качестве значения атрибута Rating. Узел Describe ChemicalEntities дает предметному посреднику команду на добавление в XML-документ с ответом описаний химических сущностей, соответствующих глобальным идентификаторам SystemID, SubstanceID и ModificationID. Описания соответствующих сущностей представляют узлы SystemXML, SubstanceXML и ModificationName. В них содержатся описания сущностей согласно данным метабазы. Фрагмент XML-документа, содержащего ответ предметного посредника на рассмотренный выше запрос приведен на рис. 5.2.11.

Как видно из XML-документа на рис. 5.2.11, предметный посредник вернул ответ, содержащий значения акустооптических свойств для арсенида галлия GaAs. Экспертный рейтинг этой информации равен единице (атрибут Rating узла «/root/row/PropertyXML/row» равен «1.0»). В документе приведены описания химической сущности в узлах, соответствующих XPath-выражениям «/root/row/PropertyXML/row/SystemXML», «/root/row/PropertyXML/row/SubstanceXML» и «/root/row/PropertyXML/row/ModificationName».

Краткие выводы

В главе получены следующие результаты:

- Выбран подход Local-as-View для обеспечения масштабируемой интеграции данных из ИС СНВМ на основе метода EII.
- Разработана логическая и физическая модель данных метабазы для метода виртуальной интеграции данных.
- Предложен механизм для извлечения наиболее достоверной информации из разнородных ИС СНВМ, основанный на экспертных оценках.
- Разработаны требования к реализации программных адаптеров интегрированной ИС СНВМ и описаны используемые форматы данных.
- Разработан и программно реализован модуль извлечения (предметный посредник) консолидированных данных интегрированной ИС СНВМ.

Глава 6
Использование интеграции приложений для информационной поддержки специалистов в области неорганической химии

6.1. Интеграция распределенных гетерогенных Web-приложений информационных систем

Как было отмечено выше, часто ИС по свойствам веществ, наряду со структурированными данными, содержат информацию в неструктурированном виде. Под неструктурированной информацией здесь понимается информация, которая не может быть структурирована в рамках схемы данных информационного источника существующей ИС. Это могут быть, например, аналитические обзоры, содержащие текстовое описание в произвольной форме, рисунки, графики, расчетные подсистемы. Как было показано, основная особенность неструктурированной информации заключается в том, что ее практически невозможно вырвать из контекста ИС, в которой она определена. Так, например, функционирование всех расчетных подсистем осуществляется в рамках исходного Web-приложения ИС. Более того, не только Web-приложение ИС является необходимым своего рода естественным интерфейсом к расчетным подпрограммам, но и сами подпрограммы рассчитаны на использование структурированных данных именно из БД своей ИС. Для того чтобы предоставить пользователю доступ ко всему многообразию неструктурированной информации, необходима интеграция распределенных гетерогенных Web-приложений ИС. По сути, необходимо выполнить EAI-интеграцию существующих Web-приложений ИС.

Интеграция Web-приложений различных ИС по свойствам веществ подразумевает, что пользователь сможет беспрепятственно переходить из Web-приложения одной информационной системы в Web-приложение другой информационной системы. Это позволит ему получить доступ

к информации, содержащейся в различных ИС и пользоваться их расчетными подсистемами.

При интеграции ИС на уровне Web-приложений (по сути, на уровне Web-интерфейсов для доступа к ИС) необходимо учитывать, что все Web-приложения информационных систем разрабатывались независимо друг от друга. Это означает, что все они используют собственные механизмы, обеспечивающие безопасность ИС, а именно, реализуют собственные системы, санкционирующие доступ к информации только со стороны авторизованных пользователей ИС. При успешной аутентификации пользователь авторизуется Web-приложением ИС и для него создается определенный контекст безопасности (зависящий от конкретного Web-приложения), в рамках которого пользователь получает доступ к ресурсам ИС. Все пользователи, которые не прошли аутентификацию, не авторизуются в ИС и, следовательно, не получают доступ к ИС со стороны Web-приложения. Таким образом, для обеспечения информационной безопасности ИС, интегрированной на уровне Web-интерфейсов, необходимо предусмотреть механизмы, обеспечивающие безопасные переходы пользователей от одного Web-приложения к другому. Очевидно, что, принимая во внимание разнородность объединяемых информационных систем, эти механизмы должны опираться на единые для всех Web-приложений стандарты. В настоящее время такими стандартами являются протокол HTTP (Hyper Text Transfer Protocol) [154] и интерфейс CGI (Common Gateway Interface) [161]. Именно на этих технологиях должен основываться механизм, обеспечивающий безопасный переход пользователя из интерфейса одной ИС в другую.

При работе с интегрированной системой необходим также механизм, который бы обеспечивал для пользователя не просто переход от одной ИС к другой, но и позволял бы ему при этом сразу же получать доступ к интересующей его информации. Другими словами, при интеграции ИС на уровне Web-приложений необходимо предусмотреть возможность просмотра релевантной (по отношению к просматриваемой пользователем системе) информации, содержащейся в других ИС. Например, пользователь, просматривая информацию о химической системе In–Sb в ИС «Диаграмма», должен иметь возможность ознакомиться с данными о пьезоэлектрических или нелинейнооптических свойствах соединения InSb из БД «Кристалл». То есть, очевидным является также то, что при построении распределенной интегрированной ИС на уровне Web-интерфейсов, необходимо обеспечить поиск релевантной информации во всех интегрируемых ИС.

Таким образом, необходим некоторый координирующий центр, который знает о том, какая информация в каких ИС хранится. То есть должна существовать центральная база данных, некоторым образом описывающая информацию, содержащуюся в интегрируемых ИС. Так приходим к понятию

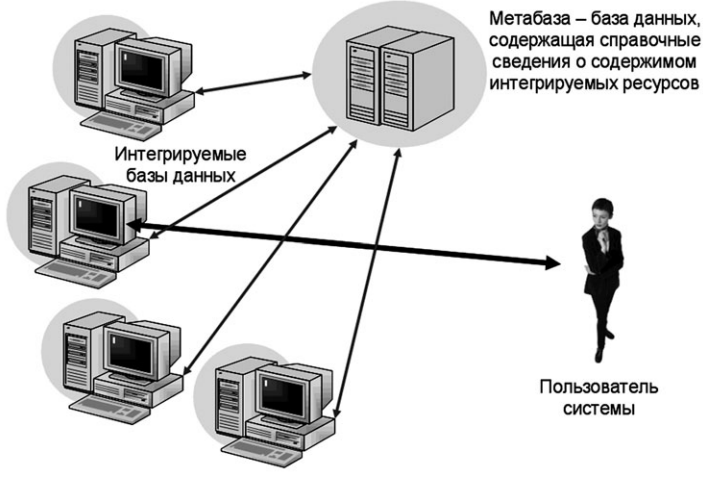


Рис. 6.1.1. Метабаза как координирующий центр интегрированной ИС

«метабаза» — специальная база данных, содержащая справочные сведения о содержимом интегрируемых ИС, а именно — о химических системах и их свойствах (рис. 6.1.1). Этих сведений достаточно для того, чтобы построить поиск релевантной информации по химическим системам и их свойствам.

Формализуем содержимое метабазы. В метабазе должна содержаться информация по интегрируемым информационным системам (множество D), по химическим веществам и системам (множество S) и по их свойствам (множество P). Для описания взаимосвязи между элементами множеств D, S и P определим тернарное отношение W на множестве $U = D \times S \times P$. Принадлежность элемента (d, s, p) отношению W, где $d \in D, s \in S, p \in P$, интерпретируется следующим образом: «В интегрируемой информационной системе d содержится информация по свойству p химической системы s».

Поиск релевантной информации по конкретной химической системе s сводится к определению отношения R, являющегося подмножеством декартова произведения $S \times S$ (иными словами, $R \subset S^2$). Таким образом, о любой паре $(s_1, s_2) \in R$ можно сказать, что система s_2 является релевантной системе s_1. Т.е., чтобы решить задачу поиска релевантной информации в интегрируемых информационных системах, необходимо некоторым образом определить отношение R. Следует отметить, что отношение R

может создаваться или дополняться компьютером по определенным правилам. Для решения этой задачи также могут быть привлечены эксперты в соответствующей предметной области (в нашем случае — химии).

Отметим, что правила построения отношения R могут быть различными. Одним из вариантов автоматического построения отношения R могут быть следующие правила:

- Для любых множеств $s_1 \in S, s_2 \in S$, состоящих из химических элементов e_{ij}, $s_1 = \{e_{11}, e_{12}, \ldots, e_{1n}\}, s_2 = \{e_{21}, e_{22}, \ldots, e_{2m}\}$ верно, что если $s_1 \subseteq s_2$ (то есть, все химические элементы из системы s_1 содержатся в системе s_2), то $(s_1, s_2) \in R$.
- Отношение R симметрично. Иными словами, для любых $s_1 \in S$, $s_2 \in S$ верно, что если $(s_1, s_2) \in R$, то и $(s_2, s_1) \in R$.

Эти два простых правила позволяют построить отношение R. Следовательно, появляется возможность определить множество систем, релевантных заданной. Как видно из этих правил, отношение R, построенное таким образом, является довольно общим, так как правила позволяют считать релевантными указанной системе все химические системы, состоящие из множества элементов, являющихся надмножеством или подмножеством указанной системы (или тем же множеством). Отметим, что такая довольно вольная трактовка релевантности не всегда является приемлемой. Часто необходимо считать релевантными указанному веществу/системе, только те вещества/системы (и, соответственно, их свойства), которые относятся к той же химической системе. Иными словами, соответствующие химические системы должны содержать тот же самый набор химических элементов. Приведем правило для автоматического построения такого отношения R:

- Для любых множеств $s_1 \in S, s_2 \in S$, состоящих из химических элементов e_{ij}, $s_1 = \{e_{11}, e_{12}, \ldots, e_{1n}\}, s_2 = \{e_{21}, e_{22}, \ldots, e_{2m}\}$ верно, что если $s_1 \equiv s_2$ (то есть, система s_1 тождественно равна системе s_2), то $(s_1, s_2) \in R$.

Заметим, что такое отношение R симметрично автоматически по определению (т. е. для любых $s_1 \in S, s_2 \in S$ верно, что если $(s_1, s_2) \in R$, то и $(s_2, s_1) \in R$).

Отметим, что здесь показаны лишь самые простые и очевидные варианты правил построения отношения релевантности R. Естественно, что ни одно определение релевантной информации не может быть универсальным и, следовательно, подходить для решения всех задач по определению релевантной информации в распределенных ИС. Поэтому на практике целесо-

образным является задание нескольких различных отношений релевантности R, которые будем называть классами релевантности. Соответственно, производить поиск релевантной информации в ИС в различных случаях можно будет с использованием разных классов релевантности.

На практике может возникнуть необходимость использования и более сложных механизмов для определения релевантной информации. Например, само понятие релевантности можно расширить и на свойства, описываемые в ИС, и даже на сами ИС. То есть, просматривая информацию по конкретному свойству химического вещества или системы в одной из интегрируемых ИС, фактически имеем информацию, определяемую тройкой (d_1, s_1, p_1). Соответственно, для этой тройки под релевантной информацией можно понимать другую тройку (d_2, s_2, p_2) — информацию по некоторому свойству химической системы из другой ИС. Следовательно, в данном случае получаем более сложное отношение релевантности, которое на этот раз будет подмножеством декартова произведения троек:

$$R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2), \text{ где } d_1, d_2 \in D; s_1, s_2 \in S; p_1, p_2 \in P.$$

Соответственно, правила для построения такого рода отношений релевантности могут учитывать не только набор химических элементов, входящих в состав соответствующих систем (как было показано выше), но и свойства, освещаемые в разных информационных системах и сами информационные системы.

Улучшение релевантности поиска можно добиться также за счет использования обозначений веществ c_i или кристаллических модификаций m_i вместо обозначений химических систем s_i в случаях, когда пользователь запрашивает релевантную информацию, находясь на уровне химических веществ или их модификаций в предложенной иерархии химических понятий.

При поиске на уровне веществ учитывается количественный состав соединения. Обозначим парой $(a_{i\min}, a_{i\max})$ количественное вхождение химического элемента $e_i \in s$ в состав, $a_{i\min}, a_{i\max} \in R^+$, $a_{i\min} \le a_{i\max}$. Если $a_{i\min} = a_{i\max}$, то вещество имеет постоянный состав по элементу $e_i \in s$. Для каждого элемента химической системы $e_i \in s$ пользователь при поиске может задать пару $(r_{i\min}, r_{i\max})$, где $r_{i\min}, r_{i\max} \in R^+$, обозначающую допустимый интервал вхождения *i*-го элемента в состав вещества (R^+ — множество неотрицательных действительных чисел). Тогда релевантными будут все вещества, относящиеся к той же химической системе, у которых для каждой пары $(r_{i\min}, r_{i\max})$ выполняется $a_{i\min} \in [r_{i\min}, r_{i\max}]$

или $a_{i\max} \in [r_{i\min}, r_{i\max}]$. Другими словами, если логическая дизъюнкция $[r_{i\min} \le a_{i\min} \,\&\, a_{i\min} \le r_{i\max}] \;+\; [r_{i\min} \le a_{i\max} \,\&\, a_{i\max} \le r_{i\max}] = true$ для всех $e_i \in s$, то данные о веществе являются релевантными [50].

При поиске релевантной информации с учетом кристаллических модификаций m_i учитываются сингонии, т. к. часто информация о кристаллических структурах может указываться по-разному. Например, для ниобата лития ($LiNbO_3$) в разных информационных источниках ИС СНВМ указывается гексагональная или тригональная кристаллическая система, что соответствует одной сингонии (гексагональной).

Как уже было отмечено, задачей интеграции ИС на уровне Web-приложений различных информационных систем является обеспечение пользователей возможностью переходить из Web-приложения одной информационной системы в Web-приложение другой информационной системы для просмотра релевантной информации. Таким образом, интеграция информационных ресурсов по свойствам веществ заключается в том, чтобы при помощи некоторого дополнительного программного обеспечения должны объединяться уже существующие Web-приложения интегрируемых ИС. Учитывая необходимость обеспечения информационной безопасности на уровне интегрированной ИС с сохранением функционирования систем безопасности интегрируемых Web-приложений, предлагается следующая схема построения интегрированной ИС на основе метабазы (рис. 6.1.2).

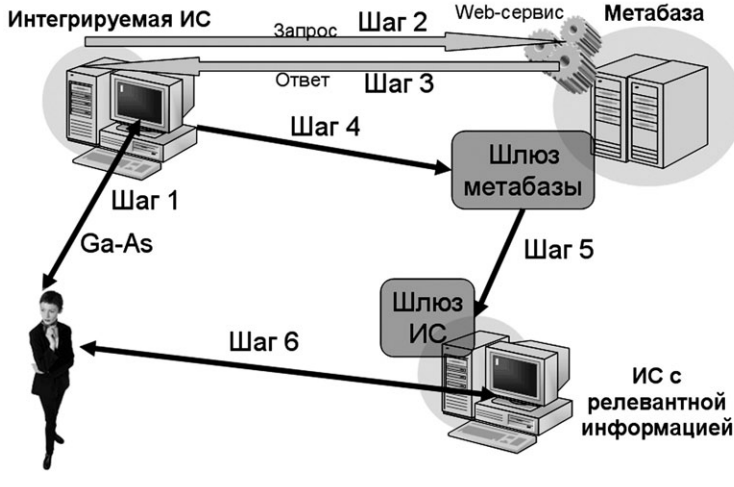


Рис. 6.1.2. Интеграция Web-приложений ИС с поиском релевантной информации в метабазе

Чтобы не перегружать рисунок, на нем приведены только две ИС, хотя на практике их будет больше, но это принципиально не меняет схему взаимодействия интегрируемых информационных ресурсов.

Рассмотрим принципы функционирования интегрированной информационной системы на базе Web-приложений с использованием метабазы для поиска релевантной информации. При создании интегрированной ИС особое внимание должно уделяться разработке ее системы безопасности. Очевидно, что у каждой ИС, подлежащей интеграции, есть собственные средства защиты информации, санкционирующие доступ к ИС. Система безопасности каждой ИС отвечает за предоставление информации только зарегистрированным пользователям данной ИС. В рамках же интегрированной системы авторизованные пользователи должны иметь право осуществлять доступ к интегрированным ресурсам строго в рамках своих прав.

Например, пользователь, который имеет доступ к ИС по фазовым диаграммам полупроводниковых систем «Диаграмма», просматривая информацию о системе In–Sb, может просмотреть данные об упругих постоянных антимонида индия в ИС «Кристалл», но не сможет посмотреть информацию о других соединениях из последней ИС, т. к. он не является зарегистрированным пользователем этой ИС или ему не разрешен полный доступ к данному ресурсу. Этот подход является общепринятым и должен быть заложен в основу распределенной системы безопасности. Для организации распределенной системы безопасности и ее совместного функционирования вместе с системами безопасности интегрируемых ресурсов, в метабазу должны пересылаться учетные данные пользователей интегрируемых ресурсов и их права доступа. Стоит отметить, что в метабазу должны передаваться не открытые пароли пользователей (это поставит под угрозу безопасность соответствующих ИС), а лишь значения хеш-функций паролей. Это позволит, с одной стороны, выполнить аутентификацию пользователя в интегрированной системе, но, с другой стороны, исключит возможность использования самих учетных данных для непосредственного входа в ИС интегрируемых ресурсов.

Поясним это на примере, приведенном на рис. 6.1.2. Пусть пользователь успешно авторизовался в ИС, т. е. выполнил вход в Web-приложение соответствующей ИС и просматривает информацию о каком-либо веществе или системе (шаг 1 на рис. 6.1.2). У каждой интегрируемой ИС есть возможность доступа к сервисному ПО, функционирующему на стороне сервера метабазы. Это ПО выполняет следующие основные функции:

- осуществляет по запросу из интегрируемых ИС поиск релевантной информации во всех интегрированных информационных системах согласно заданному классу релевантности;

- выполняет функции обеспечения информационной безопасности при переходе пользователей из одной интегрируемой ИС в другую интегрируемую ИС.

Соответственно, Web-приложение, в котором находится пользователь и выполняет просмотр информации, может направить в ПО метабазы запрос на выдачу информации, релевантной той, которую просматривает в данный момент пользователь (шаг 2 на рис. 6.1.2). Отметим, что при этом поиск релевантной информации осуществляется исходя из запрашиваемых классов релевантности. Таким образом, появляется возможность выполнять поиск по разным классам релевантности в разных случаях. Поиск также происходит с учетом прав данного пользователя на доступ к информации из других интегрируемых ИС.

В качестве ответа на запрос о поиске релевантной информации ИС метабазы должна возвращать документ, в котором будут содержаться результаты обработки запроса данного пользователя (шаг 3 на рис. 6.1.2). Так как необходимо обеспечить формат данных, который бы мог быть понят на различных программно-аппаратных платформах, предлагается использовать XML. Таким образом, в XML-документ будут помещены данные о том, какая информация по релевантным системам и их свойствам содержится в других ресурсах интегрированной ИС. Стоит отметить, что информация из XML-документа может быть выведена в любом удобном для пользователя виде с помощью XSLT-преобразования [162].

В нашем случае, целесообразно выводить релевантную информацию в виде гиперссылок, переходя по которым, пользователь смог бы осуществить безопасный переход в Web-приложение другой ИС. Однако следует помнить, что исходная ИС изначально ничего не знает о существовании других ИС. Следует отметить, что также необходимо согласование контекстов безопасности при переходе между разными ИС. По этим причинам непосредственный переход в целевую ИС из исходной ИС невозможен. Необходим некий программный шлюз, который бы выполнял функции диспетчера безопасности и санкционировал все переходы между интегрируемыми информационными системами. Таким образом, приходим к необходимости шлюза безопасности метабазы, как связующего звена при переходах пользователей между интегрируемыми ИС. Соответственно, пользователь, переходя по ссылке на релевантную информацию, должен сначала попадать на шлюз метабазы (шаг 4 на рис. 6.1.2).

В функции шлюза метабазы входит осуществление проверки подлинности пользователя, т. е. шлюз идентифицирует пользователя и проверяет, имеет ли он право обращаться к запрашиваемой информации из интегрируемой ИС. Если пользователь имеет право осуществлять доступ к запрашиваемой информации, то шлюз должен переадресовать его на специальный

входной шлюз искомого Web-приложения (шаг 5), с которого осуществляется вход в целевую ИС пользователей интегрированной информационной системы. На шлюзе ИС должен проверяться факт переадресации пользователя со шлюза метабазы и, в случае успеха, создаваться необходимый контекст безопасности для работы перешедшего пользователя в новой ИС с заданными правами доступа. После этого пользователь должен быть автоматически перенаправлен на страницу целевого Web-приложения, содержащую запрашиваемую релевантную информацию. С этого момента пользователь фактически является пользователем Web-приложения целевой ИС, и осуществляет работу в его контексте (шаг 6 на рис. 6.1.2).

Следует отметить, что, несмотря на кажущуюся сложность описанного процесса перехода пользователя из контекста одного Web-приложения в другое, данный процесс должен происходить для него прозрачно, конечно, при условии, что он имеет право на доступ к целевой ИС.

В главе 3 был предложен комплексный подход к созданию интегрированной информационной системы (ИС). В рамках этого подхода разработаны две подсистемы, которые позволили реализовать ИС, нацеленную как на конечного пользователя, так и на взаимодействие другими программными средами.

Реализация первой подсистемы в рамках комплексного подхода к созданию интегрированной ИС подразумевала применение подхода EAI для интеграции Web-приложений разных ИС по свойствам неорганических веществ. Это позволило предоставить конечному пользователю доступ ко всем информационным ресурсам интегрируемых ИС, доступным через Интернет (расчетные подсистемы и функционирующие Web-оболочки БД). То есть задача интеграции существующих Web-приложений интегрируемых ИС заключалась в объединении уже функционирующих пользовательских интерфейсов с возможностью поиска релевантной информации в интегрированных ИС и обеспечении общего контекста безопасности при переходе пользователей из Web-приложения одной ИС в Web-приложение другой ИС [169].

Реализация второй подсистемы в рамках комплексного подхода к созданию интегрированной ИС подразумевала интеграцию источников информации ИС по свойствам неорганических веществ, что соответствует применению подхода EII. Объединение источников информации позволило получить унифицированный доступ ко всем данным, содержащимся в интегрируемых источниках информации. Это дало возможность использовать информацию, содержащуюся в интегрированной ИС, в системах поддержки принятия решений.

Совместное применение подходов EAI и EII при создании интегрированной ИС по свойствам неорганических веществ позволило не только повысить качество информационного обслуживания специалистов в области

химии, но и дало возможность использовать интегрированную информацию программами компьютерной обработки информации и системами поддержки принятия решений.

6.2. Реализация интеграции гетерогенных Web-приложений информационных систем

В разделе 3.5 рассматривалась возможность построения ИС на базе доминирующих в настоящий момент технологических Web-платформ Windows Server + IIS и Unix +Apache. В результате оценки платформ согласно выбранным критериям было принято решение использовать технологические платформы компании Microsoft (Windows Server + IIS).

6.2.1. Описание структуры метабазы

В настоящее время все современные ИС, так или иначе, имеют дело с накоплением и обработкой информации. Для хранения больших объемов данных и обеспечения быстрого доступа к ним широко используются базы данных. При построении интегрированной ИС, объединяющей Web-интерфейсы ИС по свойствам неорганических веществ, необходимо обеспечить хранение информации о содержимом интегрируемых ИС, т. е. метаданных. Метаданные только некоторым образом описывают информацию, подлежащую консолидации [63]. Эти данные необходимы для обеспечения поиска релевантной информации согласно методике, описанной в разделе 6.1. Для хранения данных (метаданных) было принято решение использовать реляционную БД, именуемую в дальнейшем метабазой, т. к. она содержит справочную информацию по содержимому интегрируемых ИС. Для управления информационной БД, лежащей в основе интегрированной ИС, было принято решение использовать Microsoft SQL Server 2008.

Структура метабазы во многом определяет функциональные возможности интегрированной ИС, поэтому в настоящей главе повышенное внимание будет уделено рассмотрению структурных деталей БД, лежащей в основе ИС по объединению Web-приложений интегрируемых ИС [169]. Для решения задачи объединения Web-приложений ИС СНВМ предлагается следующая структура метабазы (рис. 6.2.1) [172].

Рассмотрим кратко назначение таблиц метабазы для интеграции Web-ресурсов по свойствам веществ, указывая, какую функциональность, опирающуюся на структуру этих таблиц, будет реализовывать интегрированная ИС.

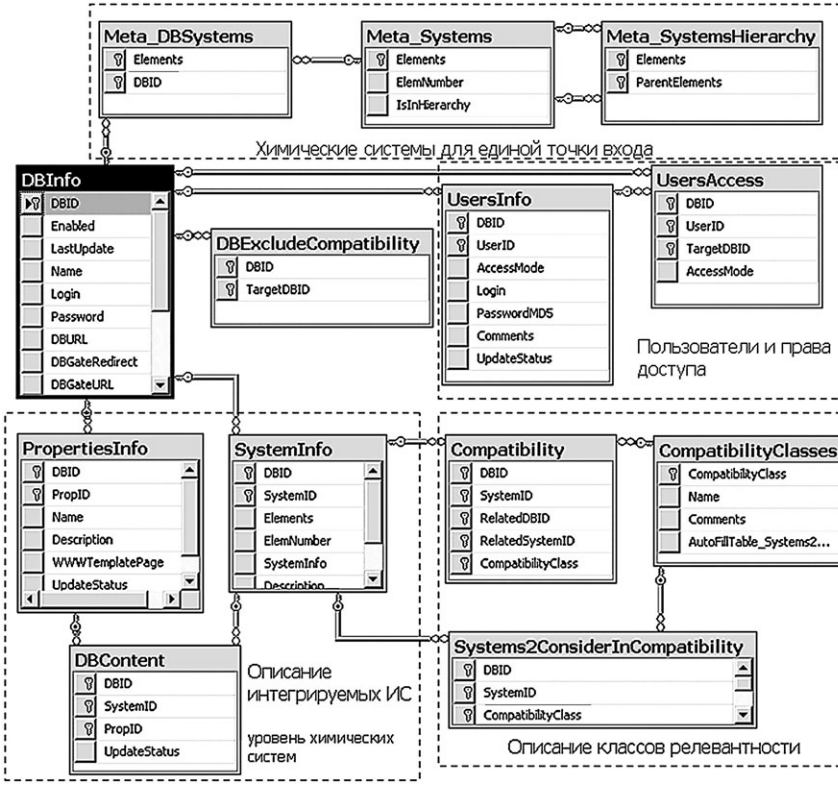


Рис. 6.2.1. Структура метабазы информационной системы, интегрирующей Web-приложения информационных систем по свойствам неорганических веществ

Таблица DBInfo

Является главной таблицей, в которой хранится список ИС по свойствам неорганических веществ, которые подключены к интегрированной на уровне Web-интерфейсов ИС. Каждой подключаемой ИС присваивается уникальный целочисленный идентификатор DBID (тип int), который является первичным ключом таблицы DBInfo. Таким образом, каждая интегрируемая ИС однозначно идентифицируется по полю DBID. Можно сказать, что таблица DBInfo задает множество D, описанное в разделе 6.1. В данной таблице также содержится информация, кратко описывающая данные информационные ресурсы и необходимая для их сопряжения с интегрированной ИС.

Приведем назначение некоторых полей этой таблицы. Поле Enabled (тип bit) указывает, активна ли интегрируемая ИС (1) или ее обслуживание

временно приостановлено (0). Этот механизм может использоваться для временного «отключения» интегрируемой ИС от общей интегрированной ИС, когда, например, возникают технические или политические вопросы, связанные с функционированием интегрируемой ИС. Необслуживаемая ИС не может обновлять информацию в метабазе и использовать все ее Web-сервисы (поиск релевантных систем и т. д.).

Поля Login (тип varchar(16)) и Password (тип varchar(16)) содержат имя учетной записи и пароль, которые должна использовать интегрируемая информационная система при вызове Web-сервисов интегрированной ИС. В поле DBURL (тип varchar(256)) содержится URL-адрес Web-приложения, которое обслуживает интегрируемый ресурс, например, «http://bg.imet-db.ru». Поле DBGateURL (тип varchar(256)) содержит URL-адрес Web-страницы, играющей роль шлюза безопасности интегрируемой ИС, через который осуществляется авторизованный вход пользователей в данную ИС, например, «http://bg.imet-db.ru/GateBG.asp?chk = #CHKSUM#& access = #ACCESS#». Формат записи этого поля будет разобран позже, при кратком обсуждении механизмов системы безопасности интегрированной ИС.

Поле WWWTemplatePage (тип varchar(256)) содержит шаблон адреса страницы для доступа к информации интегрируемого ресурса, например, «subst_res.asp?ids = #IDS#». Формат записи поля будет разобран позже. Следует отметить, что это поле может быть пусто в зависимости от механизмов переадресации, реализованных на стороне шлюза безопасности интегрируемой ИС, задаваемого в поле DBGateURL.

Таблица DBExcludeCompatibility

Данная таблица содержит информацию об интегрируемых ИС, которые надо исключать при поиске релевантной информации. Данный функционал необходим в том случае, когда одинаковая информация содержится в двух и более информационных источниках. Например, интегрированная ИС включает в качестве двух независимых ИС русскоязычную и англоязычную версию ИС «Кристалл», доступных по разным URL. Первичным ключом таблицы является связка полей DBID (тип int) и TargetDBID (тип int). Этот составной ключ идентифицирует переход из ИС с идентификатором DBID в ИС с идентификатором TargetDBID. Чтобы пользователь, находящийся, например, в русскоязычном приложении ИС Кристалл не получал список релевантной информации в англоязычной версии в отношение задаваемое таблицей вводится кортеж (1,6), где 1 — это DBID, указывающий на русскоязычную ИС Кристалл, а 6 — TargetDBID, указывающий на англоязычную.

Таблица UsersInfo

В ней содержится информация о пользователях интегрируемых ИС и их правах доступа. Первичным ключом таблицы является связка полей

DBID (тип int) и UserID (тип int). Этот составной ключ идентифицирует пользователя интегрируемой ИС.

Важным является поле AccessMode (тип int), которое указывает, какие права доступа по умолчанию предоставлены данному пользователю к информации, расположенной других в интегрируемых ИС. Определены следующие значения этого поля:

- 0 – не предоставлять пользователю доступ к интегрированным ИС;
- 1 – предоставлять пользователю доступ к интегрированным системам с наложением ограничений (осуществляется отображение информации только по той системе/веществу, которая была запрошена при входе в ИС);
- 2 – предоставлять пользователю полный доступ на просмотр информации, содержащейся в интегрированных ИС.

В полях Login (тип varchar(16)) и PasswordMD5 (тип varchar(32)) содержатся учетные данные пользователей интегрированной ИС. Следует отметить, что поле PasswordMD5 содержит MD5-хэш пароля (а не открытый пароль) пользователя интегрированной ИС, который нужен для его аутентификации. Это, с одной стороны, позволяет успешно выполнить аутентификацию пользователя, а с другой стороны, безопасность интегрируемых ИС не может быть скомпрометирована даже тогда, когда злоумышленнику удастся получить MD5-хэши паролей пользователей ИС.

Таблица UsersAccess

Необходима для переопределения прав доступа пользователя при обращении к конкретной интегрируемой ИС. Таким образом, есть возможность гибкой настройки прав доступа всех пользователей к конкретным ресурсам. С технической точки зрения, права переопределяются для конкретного пользователя ИС, определяемого идентификаторами [DBID, UserID], при доступе к конкретному ресурсу, определяемому полем TargetDBID (тип int, это внешний ключ к полю DBInfo.DBID). Сами права задаются с помощью поля AccessMode, которое может принимать те же значения, что и одноименное поле из таблицы UsersInfo (см. выше).

Таблица PropertiesInfo

В ней содержится информация о физико-химических свойствах, данные о которых хранятся в интегрируемых ИС. Другими словами, эта таблица задает множество свойств P, описанное в разделе 6.1. Первичным ключом таблицы является связка полей DBID (тип int) и PropID (тип int), указывающая на конкретное свойство, информация о котором содержится в интегрируемой ИС, т. е. PropID является идентификатором свойства в соответствующей ИС. Поле Name (тип varchar(256)) содержит наименование соответствующего физико-химического свойства, например, «температура

Кюри». Поле WWWTemplatePage (тип varchar(256)) содержит шаблон адреса страницы для доступа к информации интегрируемого ресурса, например, «properties/sing_enter.asp?asp = tb_param_el.asp&prop = #NAME#& nom = #IDS#&property = #IDP#». Содержимое этого поля зависит от принципов работы шлюза безопасности на стороне интегрируемой ИС и более подробно будет рассмотрено ниже.

Таблица SystemInfo

В ней хранится информация о химических системах, сведения о которых содержатся в интегрируемых ИС. Иными словами, эта таблица задает множество химических систем S, описанное в разделе 6.1. Первичным ключом таблицы является связка полей DBID (тип int) и SystemID (тип int). В поле Elements (тип varchar(32)) содержится список химических элементов, из которых состоит система, через тире. Притом данная строка должна начинаться и заканчиваться тире, например, «-Na-Co-Ge-O-». Поле ElemNumber (тип int) содержит количество химических элементов, образующих указанную систему. Для указанной выше химической системы это поле будет содержать значение 4. Поле SystemInfo (тип varchar(256)) содержит описание химической системы в произвольном текстовом или HTML-формате, например, это может быть «Na-Co-Ge-O» или «Na₂CoGeO₄», в зависимости от контекста, определяемого интегрируемой ИС. Поля этой таблицы заполняются Web-сервисом обновления метабазы при взаимодействии с интегрируемыми ИС, что гарантирует корректность и непротиворечивость данных в полях этой таблицы.

Таблица DBContent

В ней содержится детальная информация о том, о каких именно свойствах содержится информация в интегрированных ИС для каждой химической системы. При этом известным становится только сам факт наличия такой информации, а не конкретные значения физико-химических свойств. Первичным ключом таблицы является набор полей DBID (тип int), SystemID (тип int) и PropID (тип int). Следовательно, наличие конкретной записи, определяемой первичным ключом (DBID, SystemID, PropID), означает, что в интегрируемой ИС DBID для химической системы SystemID содержится информация о свойстве PropID. То есть кортеж данной таблицы соответствует тройке (d, s, p) в терминах раздела 2.5.

Таблица CompatibilityClasses

Как было указано в начале главы, при определении понятия релевантности невозможно задать универсальные правила, так как само понятие релевантности сильно зависит от используемого контекста. Поэтому в реализацию интегрированной ИС заложено понятие классов релевантности, которые и описываются в таблице CompatibilityClasses. Таким образом,

в этой таблице содержится информация о классах релевантности химических систем, определенных в метабазе. При этом можно добавлять новые классы релевантности, определяя правила для их автоматического построения, или осуществлять их построение вручную. Первичным ключом таблицы является поле CompatibilityClass (тип int), являющееся идентификатором класса релевантности. Поле AutoFillTable_Systems2ConsiderIn Compatibility_Flag (тип int) содержит признак необходимости добавления новых записей химических систем в таблицу Systems2ConsiderIn Compatibility для данного класса релевантности (0 — не добавлять; 1 — добавлять). Если Systems2ConsiderInCompatibility = 1, то при добавлении/обновлении информации о химической системе в таблице SystemInfo соответствующая запись для данного класса релевантности автоматически добавляется в таблицу Systems2ConsiderInCompatibility (с помощью триггера, определенного на таблице SystemInfo).

Таблица Compatibility

В ней содержится список релевантных химических систем по каждому классу релевантности. Первичным ключом таблицы являются поля DBID (тип int), SystemID (тип int), RelatedDBID (тип int), RelatedSystemID (тип int) и CompatibilityClass (тип int). Других полей в таблице нет, таким образом, все поля таблицы входят в первичный ключ. Каждая строка таблицы интерпретируется следующим образом: для химической системы [DBID, SystemID] релевантной, согласно классу релевантности Compatibility Class, является химическая система [RelatedDBID, RelatedSystemID]. Таким образом, эта таблица является главной при ответе на вопрос, какие химические системы в интегрируемых ИС являются релевантными заданной [DBID, SystemID] согласно классу релевантности CompatibilityClass. В терминах раздела 2.5 эта таблица задает отношения релевантности R для разных классов релевантности.

Таблица Systems2ConsiderInCompatibility

Является служебной и заполняется автоматически с помощью триггера при добавлении или изменении записей в таблице SystemInfo данными по добавленным или измененным химическим системам для тех классов релевантности, для которых поле AutoFillTable_Systems2ConsiderIn Compatibility_Flag из таблицы CompatibilityClasses равно 1. Это необходимо для того, чтобы подпрограммы при добавлении или изменении систем осуществляли инкрементальное перестроение классов релевантности, рассматривая только обновленные данные по системам, а не перестраивали весь класс релевантности полностью, так как эта операция может занимать продолжительное время. Таким образом, данная таблица необходима для оптимизации функционирования сервисов динамической перестройки классов релевантности и, тем самым, снижения нагрузки на сервер. Первичным

ключом таблицы являются поля DBID (тип int), SystemID (тип int) и CompatibilityClass (тип int), указывающие сервисам перестроения классов релевантности о необходимости рассмотрения обновленной химической системы (DBID, SystemID) при достраивании класса релевантности CompatibilityClass. Стоит также отметить, что при удалении записи из таблицы SystemInfo данные о соответствующей химической системе автоматически удаляются из таблиц Compatibility и Systems2ConsiderInCompatibility.

Таблица Meta_Systems

Данная таблица содержит список всех химических систем, доступный для единой точки входа в ИС CHBM. Первичным ключом таблицы является поле Elements (тип varchar(32)), содержащее список химических элементов, из которых состоит система, через тире. Данная строка должна начинаться и заканчиваться тире, например, «-Na-Co-Ge-O-». Поле ElemNumber (тип int) содержит количество химических элементов, образующих указанную систему (для быстрого поиска). Поле IsInHierarchy (тип bit) — является булевым признаком, указывающим на учет текущей системы в отношении, задаваемом таблицей

Таблица Meta_DBSystems

Содержит список химических систем, доступный в каждой ИС CHBM.

Таблица Meta_SystemsHierarchy

Задает иерархию химических систем, определяемых на множестве всех химических систем, доступных для единой точки входа в ИС CHBM (из Meta_Systems).

Таблица Versions

Эта таблица не отображена на рис. 6.1.3, т. к. не имеет непосредственного отношения к функционированию метабазы по интеграции Web-приложений ИС. Она используется для централизованного обновления версий программного обеспечения, затрагивающего структуру метабазы. В ней содержится информация о текущей версии структуры метабазы (в настоящее время это версия 9). Содержимое этой таблицы изменяется только специальным программным обеспечением, обновляющим структуру метабазы и программные модули, взаимодействующие с ней.

Таким образом, рассмотренные таблицы по их назначению можно условно отнести к нескольким группам:

- **DBInfo** — корневая таблица, содержащая информацию об интегрируемых ИС;
- **DBExcludeCompatibility** — таблица о парах ИС (задающих переходы между ИС), которые надо исключать при поиске релевантной информации;

- **UsersInfo, UsersAccess** — таблицы, содержащие информацию о пользователях интегрированных ИС и их правах доступа к другим интегрированным ресурсам;
- **SystemInfo, PropertiesInfo, DBContent** — таблицы, в которых описывается содержимое интегрируемых ресурсов (какая информация по химическим системам и их свойствам содержится в интегрируемых ИС);
- **CompatibilityClasses, Compatibility, Systems2ConsiderInCompatibility** — таблицы, содержащие информацию о доступных в метабазе классах релевантности и определяющие релевантные химические системы. Эти таблицы также отвечают за оптимизацию автоматического построения классов релевантности;
- **Meta_DBSystems, Meta_Systems, Meta_SystemsHierarchy** — нормализованные представления химических систем для быстрого поиска из ИС единой точки доступа;
- **Versions** — служебная таблица для организации системы версий ПО.

6.2.2. Загрузка информации в метабазу

В метабазе хранятся справочные данные о содержимом интегрируемых ИС, так как разные классы релевантности строятся путем заполнения таблицы Compatibility исходя именно из этой информации согласно правилам, описанным в разделе 6.1. Учитывая многообразие современных программно-аппартных платформ и трудности, возникающие при их сопряжении, выше отмечалось, что для обеспечения возможности межплатформенного взаимодействия должны использоваться открытые стандарты сетевого взаимодействия, поддерживаемые на множестве платформ. В настоящее время таким связующим звеном между различными платформами являются Web-сервисы, которые основаны на общепринятых стандартах, таких как SOAP (Simple Object Access Protocol) и XML (eXtensible Markup Language). На сегодняшний день эти технологии способны обеспечить надежную инфраструктуру для кроссплатформенного обмена сообщениями.

Таким образом, загрузка справочной информации в метабазу была реализована через Web-сервис обновления метабазы (рис. 6.2.2). Разработка велась в среде Microsoft Visual Studio 2003. Рассмотрим кратко механизм обновления метаданных (полностью данный механизм описан в [169]). Интегрируемая ИС формирует XML-документ, содержащий информацию об обновлениях данных в интегрируемой ИС. Формат этого XML-документа един для всех интегрируемых подсистем и жестко фиксирован с помощью специально разработанной XML-schema [165, 173]. Таким образом, интегрируемые подсистемы для оповещения метабазы об информационных

изменениях, произошедших в их состоянии, должны сформировать корректный XML-документ и направить его для обработки Web-сервису обновления метабазы. Взаимодействие с этим Web-сервисом осуществляется по протоколу SOAP over HTTP согласно WSDL-описанию сервиса [174]. Так в метабазе появляется актуальная информация о содержимом интегрируемых информационных ресурсов.

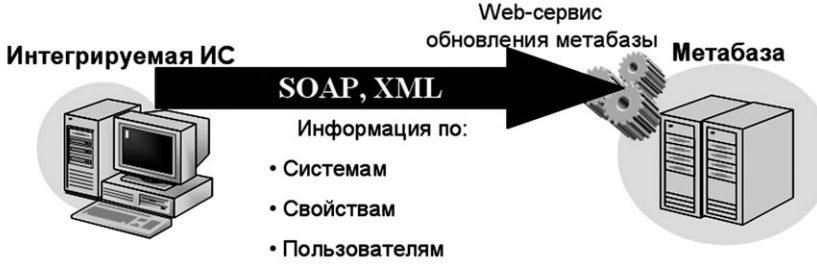


Рис. 6.2.2. Загрузка справочной информации в метабазу через Web-сервис

Для обеспечения безопасной передачи справочных данных Web-сервису обновления метабазы реализован механизм симметричного шифрования по стандарту DES (Data Encryption Standard) [175]. Дополнительно реализована возможность упаковки данных, отправляемых на сервер, для чего используется модификация zip-сжатия. Это позволяет существенно снизить объем данных, передаваемых по сети (учитывая высокую степень сжатия XML-документов), и, следовательно, понизить требования к сетям передачи данных. Такое решение является особенно актуальным в России, так как высокоскоростной доступ к сети Интернет налажен далеко не повсеместно, а механизмы упаковки сокращают объем данных настолько, что становится возможным применение обычного модема для передачи данных.

Для упрощения взаимодействия с Web-сервисом обновления метабазы был создан Web-клиент для данного сервиса. Этот программный модуль реализован в виде управляемого класса, находящегося внутри .Net сборки, функциональность которого также доступна через COM-interop. Таким образом, данный объект и доступен из любых сред, поддерживающих COM и .Net. Созданный клиент решает задачи, связанные с упаковкой и шифрованием информации, предназначенной для Web-сервиса обновления метабазы, а также управляет всеми аспектами сетевого взаимодействия, что облегчает подсоединение очередной ИС к интегрированной ИС.

После каждого сеанса обновления метабазы запускается обновление или реиндексация списка релевантных систем с учетом внесенных изме-

нений. Это позволяет поддерживать актуальную информацию о релевантных системах, содержащихся в интегрируемых ресурсах. Реиндексация разных классов релевантных систем осуществляется по предложенным в данной работе правилам поиска релевантных систем, изложенным в разделе 6.1. В случае необходимости эти правила можно изменить без кардинальной перестройки идеологии всей системы — достаточно лишь заменить программный модуль поиска релевантных систем на новый. Одним из преимуществ разработанной системы является расширяемость набора классов релевантности, что позволяет при необходимости динамически добавлять необходимые классы релевантности.

6.2.3. Поиск релевантной информации по содержимому метабазы

При интеграции Web-приложений ИС по свойствам веществ необходим поиск релевантной информации. Это избавляет пользователя от ручного поиска релевантной информации, разбросанной по различным ИС, и позволяет ему не только сэкономить время на поиск (а значит, делает его работу более продуктивной), но и гарантирует, что он не пропустит важную информацию, предоставляемую другой ИС.

Учитывая то, что средствами поиска должны пользоваться все интегрируемые Web-приложения ИС, было принято решение предоставить поисковую функциональность в виде Web-сервиса, доступного по протоколу SOAP over HTTP по URL-адресу http://meta.imet-db.ru/Service/Service.asmx. Указанный Web-сервис используется для поиска релевантной информации в ИС интегрированного комплекса и расположен на Web-сервере, обслуживающем интегрированную ИС. Данный Web-сервис служит точкой входа для интегрируемых информационных ресурсов и отвечает за предоставление им информации о релевантных системах согласно данным, содержащимся в метабазе (рис. 6.2.3) [167, 168].

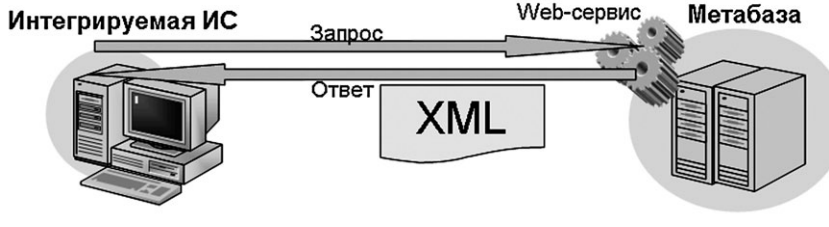


Рис. 6.2.3. Запрос релевантной информации у Web-сервиса метабазы

```
<?xml version="1.0" encoding="windows-1251" ?>
- <MetaBase date="2005-12-20T13:13:21" version="1.1">
  - <IntegratedInfo Database="crystal" SystemID="82" UserID="1" CompatibilityClass="2">
    - <DB DBID="2" DBName="Диаграмма" DBURL="http://diag.imet-db.ru">
      - <System SystemID="3" SystemInfo="As-Ga" SystemDesc="" AccessMode="1" Link="http://meta.imet-
          db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=2&tids=3&tidp=0&dt=2005-12-20T13:13:21&chk=#CHK#">
          <Property PropertyID="1" PropertyName="Аналитические обзоры" PropertyDesc="" Link="http://meta.imet-
            db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=2&tids=3&tidp=1&dt=2005-12-20T13:13:21&chk=#CHK#" />
          <Property PropertyID="5" PropertyName="Фазовые диаграммы" PropertyDesc="" Link="http://meta.imet-
            db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=2&tids=3&tidp=5&dt=2005-12-20T13:13:21&chk=#CHK#" />
          <Property PropertyID="6" PropertyName="Литературные ссылки" PropertyDesc="" Link="http://meta.imet-
            db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=2&tids=3&tidp=6&dt=2005-12-20T13:13:21&chk=#CHK#" />
          <Property PropertyID="7" PropertyName="Экспериментальные данные - точки фазовой диаграммы"
            PropertyDesc="" Link="http://meta.imet-db.ru/Gate/Gate.aspx?
            idd=1&idu=1&tidd=2&tids=3&tidp=7&dt=2005-12-20T13:13:21&chk=#CHK#" />
          <Property PropertyID="10" PropertyName="Рассчитанные данные - точки фазовой диаграммы" PropertyDesc=""
            Link="http://meta.imet-db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=2&tids=3&tidp=10&dt=2005-12-
            20T13:13:21&chk=#CHK#" />
          <Property PropertyID="32" PropertyName="Уровень качества данных о системе" PropertyDesc=""
            Link="http://meta.imet-db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=2&tids=3&tidp=32&dt=2005-12-
            20T13:13:21&chk=#CHK#" />
        </System>
      </DB>
    - <DB DBID="3" DBName="БД "Ширина запрещенной зоны"" DBURL="http://bg.imet-db.ru/">
      - <System SystemID="137" SystemInfo="GaAs" SystemDesc="" AccessMode="1" Link="http://meta.imet-
          db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=3&tids=137&tidp=0&dt=2005-12-20T13:13:21&chk=#CHK#">
          <Property PropertyID="1" PropertyName="Ширина запрещенной зоны" PropertyDesc="" Link="http://meta.imet-
            db.ru/Gate/Gate.aspx?idd=1&idu=1&tidd=3&tids=137&tidp=1&dt=2005-12-
            20T13:13:21&chk=#CHK#" />
        </System>
      </DB>
    </IntegratedInfo>
  </MetaBase>
```

Рис. 6.2.4. Пример XML-документа, возвращаемого Web-сервисом поиска релевантной информации (снимок экрана из Microsoft IE)

Взаимодействие с Web-сервисом осуществляется по протоколу SOAP согласно WSDL-описанию сервиса [176]. Для упрощения взаимодействия с Web-сервисом был также создан Web-клиент — специализированный программный модуль. Он реализован в виде управляемого класса, находящегося внутри .Net сборки, функциональность которого также доступна через COM-interop. Таким образом, данный программный модуль доступен из любых сред, поддерживающих COM и .Net. Созданный клиент решает задачи, связанные с управлением всеми аспектами сетевого взаимодействия, что облегчает подсоединение очередной ИС к интегрированной ИС. Все вопросы, связанные с Web-сервисом поиска релевантной информации, вариантами его использования и компонентом для взаимодействия с ним, подробно освещены в соответствующей документации [169].

Основной функцией, поддерживаемой Web-сервисом поиска релевантной информации, является предоставление списка релевантной информации. Данная функциональность поддерживается методом GetIntegrated Info4UserByCompatibilityClass, возвращающим для указанного пользователя данные о том, какая информация доступна (с учетом его прав) для данного пользователя в интегрируемых ИС, учитывая запрошенный класс релевантности. Информация о релевантных химических системах, содержащаяся в других информационных системах, возвращается в XML-документе. Формат этого XML-документа жестко фиксирован с помощью специально разработанной XML-schema [165, 177]. Пример XML-документа, возвращаемого при запросе пользователем (с идентификатором UserID = 1 в ИС «Кристалл») релевантной информации к химической системе с идентификатором SystemID = 82 согласно классу релевантности с идентификатором 2, приведен на рис. 6.2.4. Формат этого XML-документа будет подробнее разобран ниже при описании механизма безопасного перехода пользователя из одного Web-приложения в другое.

6.2.4. Осуществление безопасного перехода пользователя между Web-приложениями интегрируемых информационных систем

Наличие Web-сервиса поиска релевантной информации в интегрированной ИС позволяет любому пользователю получить список релевантной информации, содержащейся в ИС, о существовании которых пользователь может и не знать. Более того, даже если пользователь знает о существовании ИС, ему для входа в нее необходимо иметь статус зарегистрированного пользователя, что потребует ввода учетного имени и пароля, санкционирующего доступ к соответствующей ИС. Но даже после ввода учетных

данных пользователь должен будет самостоятельно найти интересующую его информацию, воспользовавшись навигацией по ИС.

Для обеспечения прозрачного и безопасного перехода пользователя из контекста одного Web-приложения в контекст другого Web-приложения используется схема, проиллюстрированная на рис. 6.2.5. Рассмотрим пошагово процесс перехода пользователя.

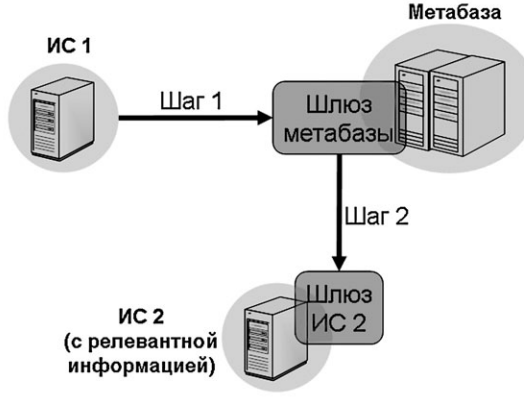


Рис. 6.2.5. Процесс перехода пользователя между Web-приложениями информационных систем через шлюз метабазы

Сначала пользователь запрашивает список релевантной информации у Web-сервиса поиска релевантной информации и получает в ответ XML-документ (см. рис. 6.2.1), содержащий набор шаблонов гиперссылок на релевантную информацию. Пример такого шаблона: «http://meta.imet-db.ru/Gate/Gate.aspx?idd = 3&idu = 1&tidd = 1&tids = 82&tidp = 1&dt = 2005-12-22T16:40:51&chk = #CHK#». После этого ИС, которая получила такой XML-документ, выполняет замену подстроки «#CHK#» на строку, вычисленную следующим образом:

НИЖНИЙ_РЕГИСТР(MD5_ПАРОЛЬ_ПОЛЬЗОВАТЕЛЯ +

"&idd = " + idd + "&idu = " + idu + "&dt = " + dt + token).

Здесь MD5_ПАРОЛЬ_ПОЛЬЗОВАТЕЛЯ — MD5-хэш пароля пользователя в ИС, из которой совершается переход. Напомним, что значения MD5-хэш функций от паролей всех пользователей ИС загружаются в метабазу

через Web-сервис обновления метабазы. НИЖНИЙ_РЕГИСТР — функция, переводящая строку в нижний регистр. В нашем случае, после замены подстроки «#CHK#» получится следующая ссылка

«http://meta.imet-db.ru/Gate/Gate.aspx?idd = 3&idu = 1&tidd = 1&tids = 82&tidp = 1&dt = 2005-12-22T16:40:51&chk = 0b0ef8086de16a095fd6b8ab48b107fa».

Отметим, что по этой ссылке осуществляется переход на шлюз безопасности метабазы. При этом передается ряд параметров (idd, idu, tidd, tids, tidp, dt, chk). Рассмотрим назначение параметров, воспринимаемых шлюзом безопасности [169]:

- **idd** — параметр, содержащий идентификатор информационной системы, пользователь которой осуществляет переход (исходная ИС или «ИС 1» на рис. 6.2.5);
- **idu** — параметр, содержащий идентификатор пользователя информационной системы, осуществляющего переход;
- **tidd** — параметр, содержащий идентификатор информационной системы, в которую пользователь запросил переход (целевая ИС или «ИС 2» на рис. 6.2.5);
- **tids** — параметр, содержащий идентификатор химической системы (в целевой информационной системе), информацию по которой запросил пользователь;
- **tidp** — параметр, содержащий идентификатор свойства (в целевой информационной системе), информацию по которому запросил пользователь;
- **dt** — параметр, содержащий дату и время формирования гиперссылки Web-сервисом системы поиска релевантной информации в формате yyyy-MM-ddTHH:mm:ss;
- **token** — параметр, содержащий маркер безопасности, получаемый от службы маркеров безопасности;
- **chk** — параметр, содержащий MD5-хеш для аутентификации пользователя и проверки параметров его запроса.

Таким образом, в сформированной ссылке содержатся параметры, определяющие время запроса, пользователя совершающего переход, целевую ИС и запрашиваемую в ней информацию. Правильно заполненный параметр chk является своего рода пропуском в запрашиваемую ИС СНВМ на просмотр релевантной информации по указанным параметрам при успешной проверке маркера безопасности. При этом следует отметить, что chk зависит от всех параметров запроса и от значения MD5-хэша пароля пользователя, поэтому нельзя изменить параметры запроса без перерасчета параметра chk. Более того, даже зная формулу для расчета

параметра chk, нельзя его рассчитать, не зная значения MD5-хэша пароля пользователя. Это позволяет пользователю осуществить безопасный переход на шлюз безопасности метабазы, щелкнув по соответствующей ссылке в Web-интерфейсе соответствующей ИС. На рис. 6.2.6 показано как может быть отображена релевантная информация в окне браузера пользователя на примере Web-приложения ИС «Bandgap».

На шлюзе безопасности метабазы происходит анализ параметров запроса [169] и их проверка. Если права доступа для данного пользователя разрешают просмотр информации в интегрируемом ресурсе, то шлюз безопасности метабазы осуществляет автоматическое перенаправление на шлюз целевой ИС (шаг 2 на рис. 6.2.5). После этого на странице шлюза целевой ИС достаточно аутентифицировать шлюз безопасности метабазы, создать для пользователя подходящий контекст безопасности и перенаправить его на страницу, указанную шлюзом безопасности метабазы в параметре RedirectToPage или на другую страницу, исходя из логики работы шлюза безопасности целевой ИС. Более детально логика типового шлюза рассмотрена в [169]. Таким образом, методика обеспечения информационной безопасности при переходе пользователя между узлами интегрированной ИС может быть схематично представлена следующей последовательностью шагов, отображенных на рис. 6.2.7.

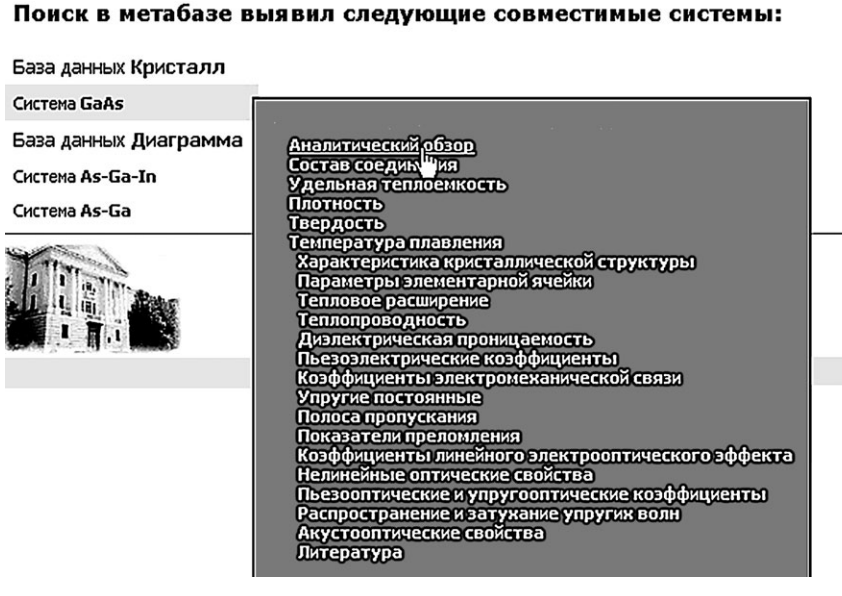


Рис. 6.2.6. Пример отображения релевантной информации в ИС «Bandgap»

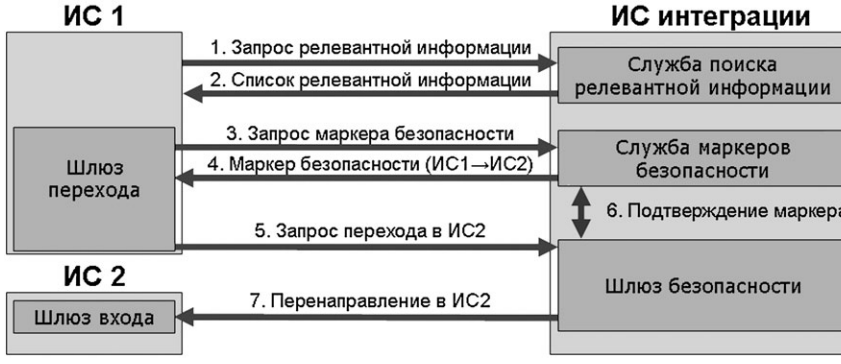


Рис. 6.2.7. Интеграция Web-приложений ИС с поиском релевантной информации

Отметим, что сама схема интегрирования по такому принципу является достаточно гибкой, так как сама политика «фильтрования» содержимого интегрируемого информационного ресурса задается настройками безопасности метабазы, а применяется на шлюзовой странице целевой ИС в контексте конкретного Web-приложения ИС. Соответственно, учитываются модели безопасности, применяемые для информационной защиты именно данного ресурса. Так, например, можно накладывать фильтр на содержимое ресурса, руководствуясь, в принципе, любыми соображениями, исходя из параметров idd, idu, tidd, tids, tidp, dt, access. Все это обеспечивает гибкость построения системы безопасности распределенной ИС СНВМ [171].

В настоящее время интеграция на уровне Web-приложений проведена для ИС «Bandgap», «Кристалл» (русско- и англоязычные версии), «Диаграмма», «Фазы», «Elements», «Кремний» и «AtomWork» (бывш. Pauling File, разработанная в NIMS, Япония) [170]. Таким образом, пользователи любой из этих ИС могут просматривать релевантную информацию, содержащуюся в смежных ИС СНВМ. Планируется функциональное расширение интегрированной ИС за счет подключения новых ИС по свойствам неорганических веществ. В частности, планируется подключить справочник по термическим константам веществ (http://www.chem.msu.su/cgi-bin/tkv.pl?show = welcom.html).

6.3. Единая точка входа в ИС СНВМ

Очень часто при поиске данных по свойству того или иного вещества неискушенный пользователь не знает к какой ИС СНВМ стоит прибегнуть для первичного сбора информации. Поэтому актуальным является создание

специализированной ИС, позволяющей потребителю данных по свойствам неорганических веществ получить возможность просмотра связанной информации по свойствам заданной химической системы в разных ИС СНВМ из одного места, которое условно называется «единой точкой входа». Созданию именно такой ИС, являющейся единой точкой входа для пользователя в ИС СНВМ, посвящен этот раздел [215].

6.3.1. Поиск релевантной информации

Основной идея заключается в предоставлении пользователю возможности выбора химических элементов, образующих химическую систему. Имея набор выбранных пользователем элементов, ИС единой точки входа должна осуществить поиск ИС СНВМ, содержащих сведения о свойствах фаз выбранной химической системы, для чего используется метабаза, разработанная ранее при создании интегрированной ИС СНВМ.

Как было описано ранее, в метабазе содержится информация по интегрируемым ИС (множество D), химическим системам (множество S) и их свойствам (множество P). Для описания взаимосвязи между элементами множеств D, S и P было определено тернарное отношение W на множестве $U = D \times S \times P$. Принадлежность элемента (d, s, p) отношению W, где $d \in D, s \in S, p \in P$, интерпретируется следующим образом: «в интегрируемой ИС d содержится информация по свойству p химической системы s».

Поиск релевантной информации s сводится к определению вида отношения R, являющегося подмножеством декартова произведения $S \times S$ (иными словами, $R \subset S^2$). Таким образом, о любой паре $(s_1, s_2) \in R$ можно сказать, что система s_2 является релевантной системе s_1. Т.е., чтобы решить задачу поиска релевантной информации в интегрируемых информационных системах, необходимо определить отношение R.

При построении ИС единой точки входа в ИС СНВМ отношение релевантности строится следующим образом: для любых множеств $s_1 \in S, s_2 \in S$, состоящих из химических элементов e_{ij}, $s_1 = \{e_{11}, e_{12}, \ldots, e_{1n}\}$, $s_2 = \{e_{21}, e_{22}, \ldots, e_{2m}\}$ верно, что если $s_1 = s_2$, то $(s_1, s_2) \in R$. Как видно из условия, отношение R симметрично. Таким образом, получим в качестве релевантных только те химические системы, вещества и модификации, которые состоят из одного и того же набора химических элементов (одинаковые химические системы). Как правило, этот способ построения отношения R является наиболее часто используемым при поиске всех свойств заданного химического вещества или системы через единую точку входа.

Поскольку поиск релевантной информации выполняется в метабазе, единая точка входа предоставляется для всех ИС СНВМ, описанных в метабазе. В настоящее время интегрированная ИС СНВМ консолидирует все разработанные в ИМЕТ РАН информационные системы: «Фазы», «Elements», «Диаграмма», «Кристалл» и «Bandgap». Благодаря проделанной работе по международной интеграции, удалось включить в состав интегрированной системы ИМЕТ РАН ИС «AtomWork» (разработанную в NIMS, Япония), содержащую информацию о более чем 23 тыс. неорганических веществ [271, 272].

6.3.2. Разработка Web-приложения ИС

Рассмотрим кратко особенности разработки Web-приложения единой точки входа, располагаемого по адресу http://meta.imet-db.ru. Web-приложение ASP.Net написано на языке C# (.Net Framework 3.5) с использованием ADO.Net для доступа к метабазе. Для построения запросов используются языковые средства Transact-SQL, являющегося диалектом языка SQL, который используется в СУБД Microsoft SQL Server 2008.

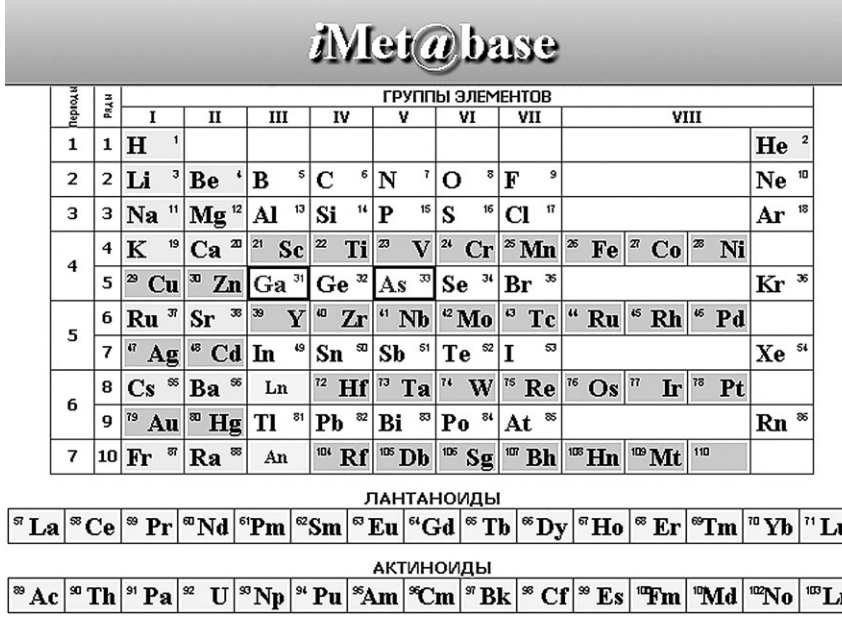


Рис. 6.3.1. Выбор химической системы по набору элементов

Пользовательский интерфейс является интерактивным за счет использования библиотеки jQuery, облегчающей взаимодействие с HTML DOM (Document Object Model — объектная модель документа) и предоставляющей удобный интерфейс (API) для работы с AJAX (Asynchronous Javascript and XML — асинхронный JavaScript и XML).

Опишем принцип работы Web-приложения. Основной элемент интерфейса пользователя — интерактивная таблица Менделеева (рис. 6.3.1). Пользователю предоставляется возможность выбора химических элементов, образующих химическую систему. При нажатии на каждый химический элемент (выбор или снятие выбора) происходит его подсветка за счет применения классов из каскадных таблиц стилей (CSS) с помощью библиотеки jQuery (язык программирования JavaScript):

```
$(".Mendeleev .element").click(function() { // клик на элементе
$(this).toggleClass("selected"); // переключаем класс
var arr = [];
$(".Mendeleev .selected").each(function() {
arr.push($(this).children(".name").text());
});
arr.sort();
var st = arr.join("-");
$(".result").html("");
if (st = = "") {
$(".Mendeleev .inactive").show();
$(".Mendeleev .active").hide();
}
else {
$(".Mendeleev .inactive").hide();
$(".Mendeleev .active").show();
ProcElements(st);
}
$(".Mendeleev .selectedSystem").html(st);
});
```

Одновременно в случае наличия выбранных химических элементов происходит вызов функции `ProcElements`, которая обеспечивает отправку асинхронного AJAX-запроса к HTTP-обработчику http://meta.imet-db.ru/ JSON_Elements.ashx, являющемуся сервисом поиска релевантной информации:

```
function ProcElements(elements) {
$(".result").html("<center><img src = '/i/loaderlight.gif'
alt = 'подождите…' width = 24 height = 24 /></center>");
$.ajax({
type: "post",
url: "/JSON_Elements.ashx",
data: { "mode": "getelementsinfo", "elements": elements },
dataType: "json",
error: function(XMLHttpRequest, textStatus, errorThrown) {
```

```
$(".result").html("<center><span class = 'err'>ajax error
textStatus = " + textStatus + ", errorThrown = " + errorThrown +
"</span></center>");
},
success: function(json) {
if (json.MsgRu ! = "" || json.MsgEn ! = "") {
$(".result").html("<center><span class = 'err ru'>" + json.MsgRu +
"</span><span class = 'err en'>" + json.MsgEn + "</span></center>")
return;
}
var st = "";
if (json.Data.Table[0].Row.length = = 0) {
st = "<center><span class = 'err ru'>нет данных</span>
<span class = 'err en'>нет данных</span></center>";
}
else {
st = RenderDataAsTable_PopUp(json); // список-попап
}
$(".result").html(st);
}
});
}
```

При выборе, например, химической системы As–Ga будет отправлен следующий POST запрос на адрес: http://meta.imet-db.ru/JSON_ Elements.ashx, содержащий данные mode = getelementsinfo&elements = As-Ga. Задача сервиса поиска релевантной информации — по множеству выбранных химических элементов вернуть перечень ИС СНВМ, содержащих сведения о заданной химический системе. Поскольку клиентская часть отработки информации реализована на JavaScript, как стандартном языке сценариев, поддерживаемом всеми браузерами, то сервис поиска формирует ответ в формате JSON (JavaScript Object Notation — нотация объектов JavaScript), который является естественным при использовании языка JavaScript. Приведем пример возвращаемого документа при поиске информации по элементу Ga (исключительно для краткости, т. к. по As–Ga размер JSON документа на порядок больше):

```
{"MsgRu":"","MsgEn":"","Data":{ "Table":[{ "Row":[ { "Col":
[ "5","31","-Ga-","1","Ga","","1","Свойства элемента","
","/elements/properties_all_given.aspx?elem = #IDS#",
"Элементы","http://phases.imet-db.ru/elements/","","post",
"http://phases.imet-db.ru/elements/GateElements.asp?chk =
#CHKSUM#&t = #TOKEN#","ru"]}, { "Col":[ "7","16027",
"-Ga-","1","Ga","","1","Information","","","AtomWork (NIMS,
Japan)","http://crystdb.nims.go.jp/index_en.html","","link",
"https://login-matnavi.nims.go.jp/sso/UI/Login?goto =
 http%3A%2F%2Fcrystdb.nims.go.jp%2Fcrystdb%2Fsearch-materials-
list%3FisVisiblePeriodicTable%3Dtrue%26condition_type%3Dchemical_
system%26need_more_type%3Dprototype_number%26condition_value%3D#
ELEMENTS_PLUSinURL#&IDToken1 = #CHKSUM#&IDToken1 = #TOKEN#","en"]}
]}]}}
```

Как видно, полученный JSON-документ возвращает в числе прочих данных ссылки для перехода на ИС CHBM с релевантной информацией. Однако ссылки нуждаются в дальнейшей обработке в частности для замены «#CHKSUM#» и «#TOKEN#», которые являются контрольной суммой (вычисленной с использованием хеш-функции MD5) с отпечатком параметров перехода и маркером безопасности (token) соответственно. За обработку и вывод пользователю информации из JSON-документа отвечает функция `RenderDataAsTable_PopUp(json)`, результат работы которой виден на рис. 6.3.2.

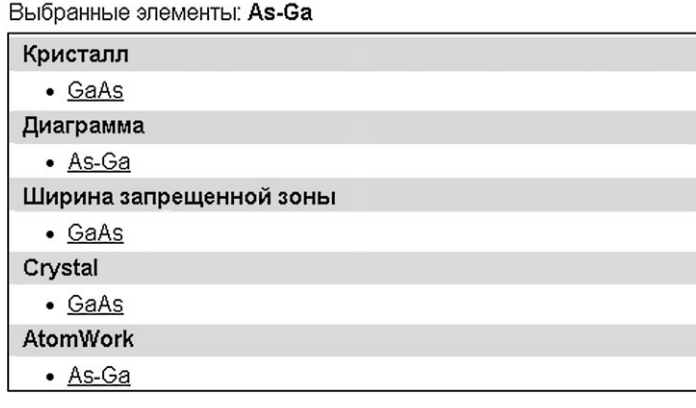


Рис. 6.3.2. Список релевантной информации в ИС CHBM для системы As–Ga

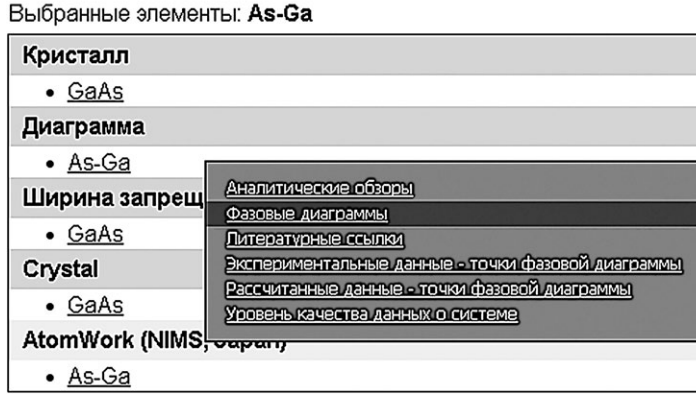


Рис. 6.3.3. Список свойств в ИС «Диаграмма» для системы As–Ga.

При наведении пользователем указателя на химическую сущность (систему, вещество или кристаллическую модификацию) выводится список свойств, доступных для просмотра в соответствующей ИС СНВМ (рис. 6.3.3).

Пользователь может, щелкнув по гиперссылке, прозрачно перейти через шлюз безопасности единой точки входа http://meta.imet-db.ru/gate/gateSAP.aspx, в ИС СНВМ с запрошенной информацией. При этом происходит автоматическое перенаправление на страницу с требуемой информацией, например при переходе по ссылке «Фазовые диаграммы» пользователь увидит сразу затребованную информацию (рис. 6.3.4).

Таким образом, в работе на основе метода интеграции информационных систем (EAI) реализована не только интеграция между гетерогенными материаловедческими ИС с безопасным переходом к просмотру релевантной информации, но и создана единая точка входа во все ИС СНВМ, описанные в каталоге информационных ресурсов метабазы (рис. 6.3.5). Белыми стрелками на рисунке показаны потоки данных при запросах релевантной информации, серыми стрелками показан переход пользователя

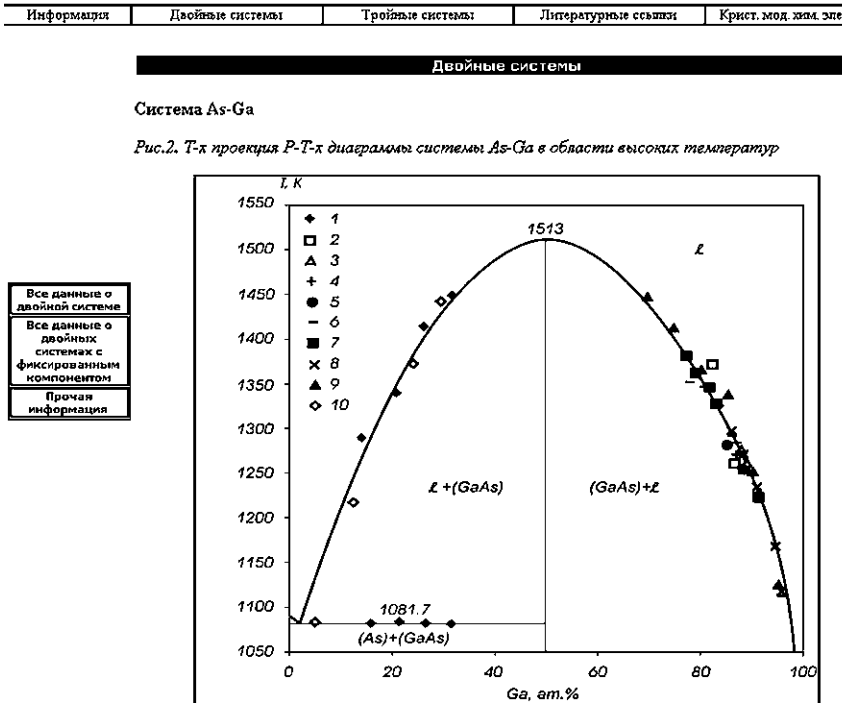


Рис. 6.3.4. Фазовая диаграмма для системы As–Ga в ИС «Диаграмма»

из единой точки входа в ИС СНВМ с релевантной информацией, а черными стрелками обозначен переход пользователя из контекста одной из ИС СНВМ в контекст ИС СНВМ с релевантной информацией.

На текущий момент разработанная интегрированная ИС СНВМ является единственной успешной попыткой интеграции материаловедческой информации на территории России. Достоверность приведенных в исследовании выводов подтверждается практической реализацией интегрированной ИС, которая может использоваться как конечными пользователями для поиска и сбора информации (EAI), так и программными средами в качестве источника информации по свойствам неорганических веществ (ETL+EII).

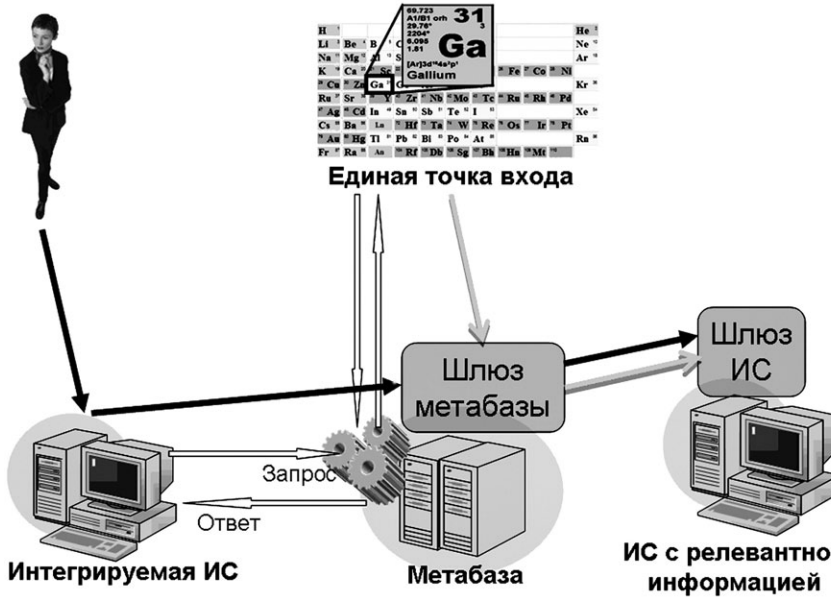


Рис. 6.3.5. Реализации интеграции ИС СНВМ методом EAI

6.4. Создание системы единой авторизации

Часто при интеграции пользовательских интерфейсов гетерогенных ИС возникает задача создания единой ситемы авторизации, известной в англоязычной литературе как Single Sign-On (SSO). Применительно к ИС СНВМ это способ обеспечения согласованной работы двух и боле Web-приложений ИС СНВМ, при использовании которого пользователь переходит от одной ИС СНВМ к другой без повторной авторизации. Говоря простым

языком — это использование единого профиля и пароля от него для входа в разные Web-приложения ИС СНВМ. К основным преимуществам технологии единого входа относятся:

- уменьшение количества запоминаемых логинов и паролей;
- уменьшение времени на повторный ввод пароля для одной и той же учетной записи;
- снижение нагрузки на IT-инфраструктуру за счет уменьшения количества запросов по восстановлению забытых паролей.

В технологии единого входа применяются централизованные серверы аутентификации, используемые другими ИС, которые обеспечивают ввод пользователем своих учетных данных только один раз для авторизации на одной из ИС.

Из недостатков, приписываемых SSO, следует выделить увеличивающуюся важность одного пароля, при получении которого злоумышленник получает доступ сразу ко всем данным, привязанным к профилю SSO. Значит, организация единого входа требует повышенного внимания к защите учетных данных пользователя. Так же утеря пароля может привести к блокировке и отказу в доступе ко всем системам, использующим SSO.

Для создания единого профиля пользователя для доступа к ИС СНВМ была разработана реляционная структура данных, представленная на рис. 6.4.1. По факту все эти реляционные таблицы являются частью схемы данных метабазы, разработанной в этой главе ранее. Названия всех таблиц, относящихся к SSO, начинаются с префикса «SSO_». Рассмотрим кратко назначения всех таблиц.

Таблица SSO_Users является основной таблицей профиля пользователя, в которой хранится список регистрационных данных пользователей, зарегистрированных в ИС. Каждому пользователю присваивается уникальный целочисленный идентификатор UserID (тип int), который является первичным ключом таблицы. Поля BusinessTypeID (тип int), CountryID (тип int), OccupationID (тип int), PostID (тип int), PurposeID (тип int), являются внешними ключами (foreign keys), для одноименных полей в таблицах-справочниках: SSO_BusinessTypes, SSO_Countries, SSO_Occupations, SSO_Posts, SSO_Purposes (тип бизнеса, страна, род деятельности, должность и цель использования). Поле AddressType (тип int) содержит тип адреса, указанного пользователем (1 — рабочий адрес; 2 — домашний адрес). Поле Org (тип nvarchar(128)) содержит название организации места работы, а поле Department (тип nvarchar(128)) содержит названия отдела организации и не обязательно для заполнения при регистрации. Поля FirstName (тип nvarchar(64)), MiddleName (тип nvarchar(64)) и LastName (тип nvarchar(64)) содержат имя, отчество и фамилию пользователя соответственно (отчество не обязательно для заполнения).

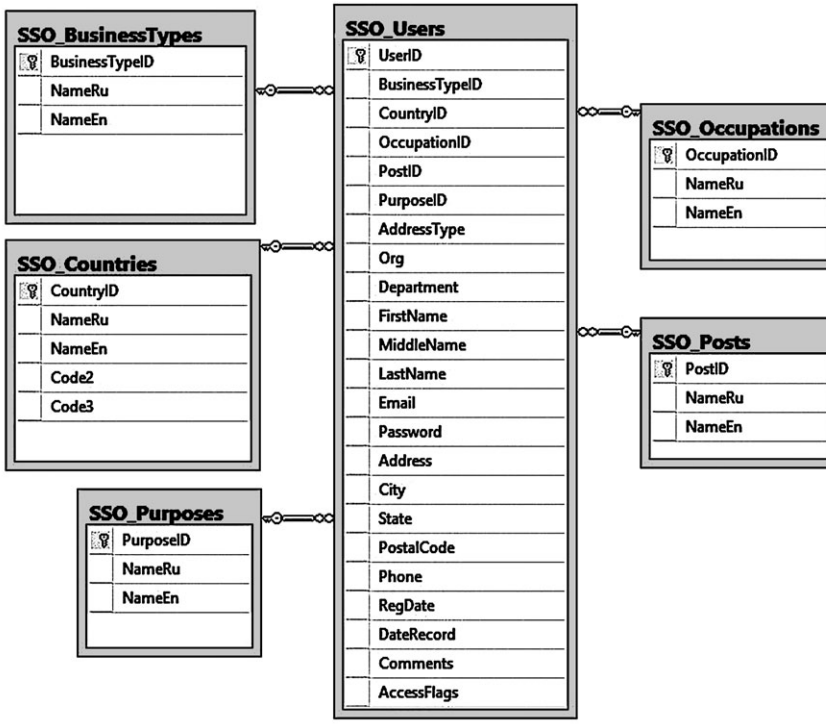


Рис. 6.4.1. Реляционная структура профиля пользователя ИС СНВМ

Поле Email (тип nvarchar(128)) содержит адрес электронной почты пользователя интегрированной ИС, и также одновременно является логином при входе. Построен уникальный индекс, содержащий это поле, что гарантирует уникальность электронных адресов (т. е. не может быть зарегистрировано более одного пользователя с одинаковыми адресами). Поле Password (тип nvarchar(32)) содержит пароль необходимый для авторизованного входа пользователей в ИС.

Поля Address (тип nvarchar(128)), City (тип nvarchar(32)), State (тип nvarchar(64)), PostalCode (тип nvarchar(16)), Phone (тип nvarchar(64)) содержат данные по адресу пользователя с почтовым индексом и телефоном. Поле RegDate (datetime) содержит дату и время регистрации, а поле DateRecord (datetime) содержит дату и время последнего обновления профиля пользователя. Поле Comments (тип nvarchar(2048)) содержит комментарии, указанные для данного пользователя администратором. Поле AccessFlags (тип int) указывает, какие права доступа по умолчанию

предоставлены данному пользователю. Определены следующие значения этого поля:

- 0 — не предоставлять пользователю доступ к полнотекстовым статьям (литературным ссылкам);
- 1 — предоставлять пользователю полный доступ.

Таблицы SSO_BusinessTypes, SSO_Countries, SSO_Occupations, SSO_Posts, SSO_Purposes являются справочниками, определяющими доступные для выбора в профиле пользователя элементы списков. Справочники организованы единообразно (ключевое поле и название на русском и английском языке), поэтому рассмотрим только справочник стран (таблица SSO_Countries) в качестве примера. Таблица **SSO_Countries** содержит записи о всех странах мира, которые доступны пользователя для выбора при регистрации. Поле CountryID (тип int) является первичным ключом (по этому полю осуществляется связь с внешним ключом SSO_Users. CountryID) и содержит код страны по стандарту ISO 3166-1. Поле NameRu (varchar(64)) содержит название страны, записанное на русском языке, а NameEn (varchar(64)) — на английском. Поля Code2 (varchar(2)) и Code3 (varchar(3)) содержат названия стран, записанные латиницей в сокращенном виде из 2 и 3 символов соответственно по стандарту ISO 3166–1. При появлении новых данных все справочники можно легко дополнять необходимой информацией.

	CountryID	NameRu	NameEn	Code2	Code3
▶	4	Афганистан	Afghanistan	AF	AFG
	8	Албания	Albania	AL	ALB
	10	Антарктида	Antarctica	AQ	ATA
	12	Алжир	Algeria	DZ	DZA
	16	Американское...	American Samoa	AS	ASM
	20	Андорра	Andorra	AD	AND
	24	Ангола	Angola	AO	AGO
	28	Антигуа и Бар...	Antigua and Ba...	AG	ATG
	31	Азербайджан	Azerbaijan	AZ	AZE

Рис. 6.4.2. Снимок экрана с фрагментом таблицы SSO_Countries

В настоящий момент на основе разработанной реляционной структуры данных создана ИС единой авторизации, поддерживающая русскоязычный и англоязычный пользовательский интерфейс. ИС доступна по адресу http://sso.imet-db.ru/ и в настоящий момент используется для авторизации

пользователей ИС «Bandgap», содержащей данные по ширине запрещенной зоны неорганических веществ, т. к. доступ к данной ИС СНВМ в настоящий момент открыт для всех бесплатно (http://bg.imet-db.ru/ — требуется только регистрация). Поля регистрационной формы показаны на рис. 6.4.3. После регистрации пользователь имеет право доступа к ИС «Bandgap».

Важно отметить, что основной проблемой, решенной при создании единой ИС авторизации, являлась организация AJAX запросов к серверу ИС авторизации, совмещенная с передачей значений cookie между разными доменами, которая в силу безопасности запрещена.

В результате анализа способов решения данной проблемы (JSONP, проксирование клиентских запросов сервером, CORS), был выбран относительно недавно появившийся механизм CORS [309]. Cross-origin resource

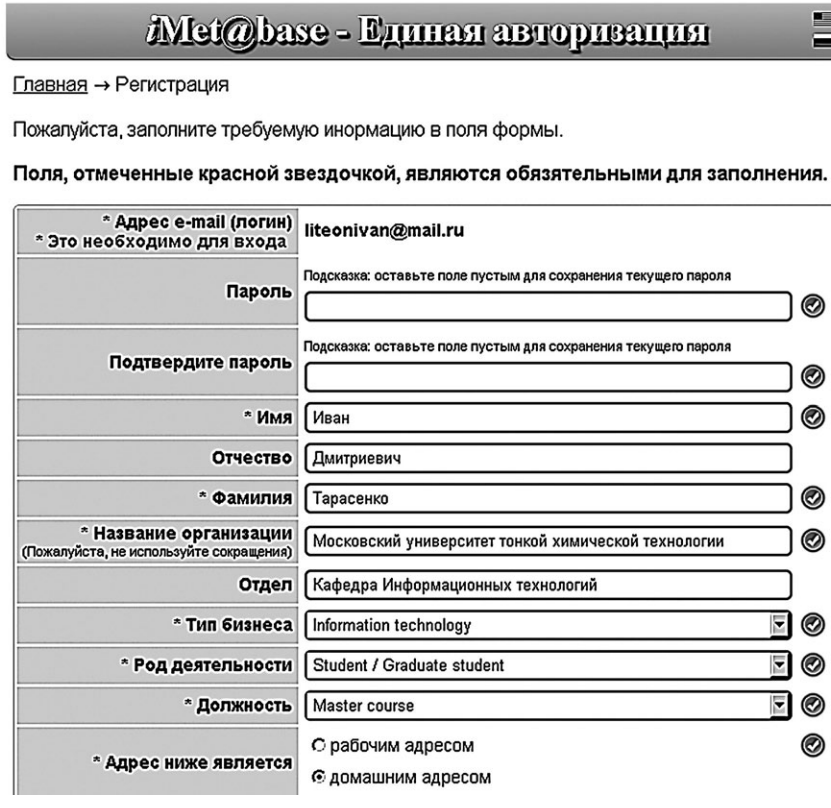


Рис. 6.4.3. Регистрация пользователя в ИС единой авторизации

sharing (CORS) — механизм, позволяющий клиентскому сценарию JavaScript создать на странице запрос к другому домену, отличному от того, с которого этот JavaScript был загружен. Подобный перекрестный между доменами запрос запрещен во всех браузерах политикой безопасности, CORS определяет механизм настройки серверных заголовков протокола HTTP таким способом, благодаря которому браузер и Web-сервер получают возможность взаимодействия. Т. е. сервер самостоятельно определяет, позволять или нет подобные перекрестные запросы со стороны других доменов с возможностью фильтрации запросов. Серверная поддержка CORS в ИС единой авторизации реализована в ASP.Net приложении следующим образом (язык C#):

```
private string GetHeaderValue4Origin {
get {
string host = Utils.SafeGetString(context.Request["host"]);
return string.IsNullOrEmpty(host) ?
"*" : string.Format("http://{0}", host);
}
}
/// <summary>
/// пишем заголовки для CORS, если допустимый хост
/// </summary>
/// <returns>true – заголовки CORS записаны</returns>
private bool WriteCORSHeaders() {
string host = Utils.SafeGetString(context.Request["host"]);
if (string.IsNullOrEmpty(host))
return false;
if (host.EndsWith(".imet-db.ru")
|| host.EndsWith(".imet.ac.ru")) {
context.Response.AddHeader("Access-Control-Allow-Origin",
GetHeaderValue4Origin); // включаем CORS
context.Response.AddHeader("Access-Control-Allow-Credentials",
"true"); // включаем CORS
return true;
}
return false;
}
```

Как видно из кода, в настоящий момент перекрестные запросы разрешены только для доменов *.imet-db.ru и *.imet.ac.ru, т. к. только с этих доменов можно совершить запрос к приложению http://sso.imet-db.ru/. Добавление заголовка Access-Control-Allow-Origin разрешает перекрестные запросы, а Access-Control-Allow-Credentials позволяет передавать значения cookie, указывающие на сессию авторизации.

Со стороны клиента авторизации, которым является ИС «Bandgap», осуществляются перекрестные AJAX-запросы на получение параметров сессии авторизованного пользователя. Важным является использование

	IE	Firefox	Chrome	Safari	Opera	iOS Safari	Opera Mini	Android Browser	Blackberry Browser	Opera Mobile	Chrome for Android	Firefox for Android	IE Mobile
3 версии назад	8.0	24.0	29.0	5.1	16.0	4.2-4.3		4.0		11.5			
2 версии назад	9.0	25.0	30.0	6.0	17.0	5.0-5.1		4.1		12.0			
Предыдущая версия	10.0	26.0	31.0	6.1	18.0	6.0-6.1		4.2-4.3	7.0	12.1			
Текущая	11.0	27.0	32.0	7.0	19.0	7.0	5.0-7.0	4.4	10.0	16.0	32.0	26.0	10.0
Ближайшее будущее		28.0	33.0		20.0								
Последующее будущее		29.0	34.0		21.0								
3 версии вперед		30.0	35.0										

Рис. 6.4.4. Поддержка CORS современными браузерами

в AJAX запросов дополнительных настроечных параметров (withCredentials и crossDomain), включающих использование CORS на стороне клиента:

```
$.ajax({
        xhrFields: { withCredentials: true },
        crossDomain: true,
…
}
```

Таким образом, применение перекрестных запросов для авторизации возможно только если одновременно клиент и сервер явно разрешают возможность их использования. На клиентской стороне поддержка CORS осуществляется всеми основными браузерами (рис. 6.4.4). Условные обозначения: белый — CORS поддерживается, черный — не поддерживается, серый — нет данных. В IE8 и IE9 CORS поддерживается ограничено, поэтому данные браузеры не могут быть использованы для работы с ИС «Bandgap», о чем выдается соответствующее сообщение.

По мере открытия свободного доступа к другим ИС СНВМ, ИС единой авторизации будет расширена и на эти системы. В будущем возможно, но маловероятно, использование ИС единой авторизации со стороны ИС СНВМ, разработанных в других организациях.

Краткие выводы

В главе получены следующие результаты:

- Разработана структура метабазы информационной системы, интегрирующей Web-интерфейсы гетерогенных ИС СНВМ и программно реализован механизм загрузки данных в нее.
- Формализовано понятие релевантной информации, содержащейся в ИС СНВМ на уровне систем, веществ и модификаций.
- Разработан и программно реализован механизм поиска релевантной информации в рамках интегрированной ИС СНВМ.
- Разработана методика обеспечения информационной безопасности при переходе пользователя между узлами интегрированных ИС СНВМ.
- Разработана единая точка входа в интегрированные ИС СНВМ (http://meta.imet-db.ru).
- Создана единая система авторизации для применения в интегрированной ИС СНВМ (http://sso.imet-db.ru).

Глава 7
Применение интегрированной информационной системы для поиска закономерностей и компьютерного конструирования новых соединений

Интеграция баз данных является первым шагом к разработке интеллектуальных информационных систем. Метабаза данных, которая содержит тезаурус профессиональных терминов, используемых в интегрируемых БД, может рассматриваться как основа интеллектуального интерфейса объединенной информационной системы. Она, в совокупности с разработанными нами прикладными программами, решает проблему поиска затребованных пользователем сведений об определенных неорганических веществах в различных БД [180, 181]. Дальнейшая интеллектуализация баз данных непосредственно связана с оснащением информационных систем программными комплексами анализа огромных массивов химической информации, которые содержит разработанная в настоящей работе интегрированная система баз данных, и с поиском закономерностей в этой информации. Найденные закономерности (знания), которые в дальнейшем будут храниться в специальной базе знаний о предметной области, позволяют сконструировать еще неполученные вещества с заданными свойствами, что расширяет возможности БД, превращая их из компьютерного справочника в интеллектуальные информационные системы [182]. Такие интеллектуальные информационные системы дают возможность прогнозировать еще экспериментально неизученные вещества, оценивать их параметры и принимать решение о путях поиска новых веществ с заданными свойствами [1].

В настоящей работе проведены поисковые исследования по использованию данных из интегрированной информационной системы для поиска сложных взаимосвязей в химической информации. Найденные взаимосвязи применены для конструирования новых неорганических соединений, перспективных для поиска новых соединений для электронной промыш-

ленности [183]. Целью проведенных исследований является демонстрация возможностей информационно-аналитической системы для компьютерного конструирования неорганических веществ [300]. Физико-химической основой разработки такой системы является периодический закон, из которого знаем, что существуют периодические зависимости между свойствами соединений и свойствами элементов, входящих в их состав. Более того, уже известные соединения, информация о которых хранится в БД, должны подчиняться этим периодическим закономерностям. Следовательно, возможен поиск таких закономерностей образования соединений определенных типов на основе анализа информации БД.

7.1. Интерполяция неизвестных значений в обучающих выборках

Одной из основных сложностей при использовании программных комплексов распознавания образов для компьютерного конструирования неорганических соединений является наличие пропусков в значениях свойств химических элементов, на основе которых формируются выборки для обучения. В зависимости от алгоритмов распознавания пропуски (отсутствия значений) могут не только исказить правильность обучения и, следовательно, распознавания, но и вызвать отказ в обучении, вынуждающий исследователя полностью отказаться от использования признака, содержащего пропуски в значениях.

Существует два варианта действий: исследовать матрицу, не заполняя пропуски, или заполнить тем или иным образом пропуски и анализировать полученную заполненную матрицу. Первый вариант более строг, однако практически всегда заставляет отказаться либо от рассмотрения значительной части фактических данных, либо от применения ряда мощных методов, «воспринимающих» только полные матрицы. Второй вариант «работает», то есть не оказывает заметного искажающего влияния на результаты анализа, когда число пропусков в данных мало (до 20 %). Рассмотрим основные приемы обработки данных, содержащих пропуски.

7.1.1. Краткий обзор методов заполнения пропусков в данных

Метод исключения неполных векторов. Метод исключения неполных (некомплектных) векторов (casewise deletion) [301] состоит в том, что все векторы (строки или столбцы матрицы), содержащие пропуски,

исключают из рассмотрения, и в дальнейшем анализируют новую, редуцированную матрицу данных. Когда выборка содержит достаточное число комплектных объектов, такой подход следует признать наиболее целесообразным.

Однако на практике распространена ситуация, когда наличие даже небольшого числа случайно (или, как минимум, без явной закономерности) распределенных пропусков — при формально большой размерности данных — приводит к резкому уменьшению числа комплектных наблюдений. Так, например, если предположить, что пропуски распределены независимо по закону Бернулли, то в случае наличия 5 % пропусков для матрицы с числом столбцов $m = 10$ ожидаемая доля комплектных наблюдений составит $0.95^{10} \sim 0.6$ (60 %), то есть редуцированная матрица будет содержать $0.6/0.95 = 0.63$ (63 %) данных, присутствующих в исходной матрице.

Метод заполнения средними значениями. Метод заполнения пропусков безусловными средними значениями (mean substitution) является одним из самых простых и известных методов заполнения пропусков. При этом пропуск заменяется средним по столбцу матрицы. При применении этого метода происходит смещение (уменьшение) дисперсии переменных, что приводит и к смещенным оценкам элементов ковариационной и корреляционной матриц. Коэффициенты ковариации оказываются занижены, а корреляции — завышены [302]. Насколько приемлемо полученное смещение для дальнейшего анализа, решается в каждом конкретном случае; в целом — метод пригоден при малом числе пропусков в данных.

Метод заполнения условными средними значениями. Метод заполнения пропусков условными средними значениями (метод Бака, imputation by regression) [302]. Для двух переменных матрицы, заметно коррелирующих между собой — m_1 и m_2, можно построить регрессионное уравнение зависимости одной переменной от другой: $m_2 = am_1 + b$ по наблюдениям, известным для обоих переменных, и оценить недостающие значения m_2 с помощью полученного регрессионного уравнения по имеющимся значениям m_1. Данные, заполненные по методу Бака, обеспечивают разумные оценки средних, в частности, если приемлемо предположение о нормальности наблюдений. Ковариационная матрица по заполненным методом Бака данным занижает величину дисперсий и ковариаций (а корреляционная матрица завышает величину корреляций), хотя и не так сильно, как при подстановке безусловных средних.

Метод заполнения выборочными значениями. Метод заполнения пропусков выборочными значениями (hot deck imputation) [302]. Существуют методы заполнения пропусков, основанные на использовании расстояния до обоих объектов (в некоторой метрике между парами объектов), которое определяется по значениям признаков. Считается, что если два

объекта близки в пространстве измеренных признаков, то из этого следует и их близость по неизмеренным признакам. Метрика и пороговое значение расстояния, определяющее близость объектов, вводятся в зависимости от условий конкретной задачи — шкал, в которых признаки измерены, количества пропусков и т. д.

Например, пусть требуется оценить значение пропущенного признака m_j, то есть оценить элемент m_{ij} матрицы. Для этого формируется подматрица с измеренными значениями признака, из которой далее выделяется группа наиболее близких объектов в пространстве измеренных у этого объекта признаков. Затем неизвестное значение m_{ij} заменяется средним по выделенной однородной группе объектов значением признака m_{ij} или случайным значением из этой группы.

Метод максимального правдоподобия. Оценка пропусков методом максимального правдоподобия (Expectation-Maximization, EM algorithm) [301]. EM-алгоритм — общий итеративный алгоритм для задач оценивания методом максимального правдоподобия и не только для заполнения отсутствующих данных — например, он используется для оценивания компонент дисперсии, итеративно взвешиваемых оценок наименьших квадратов и т. д. Достоинством EM-алгоритма является его надежная сходимость, недостатком — то, что скорость сходимости может быть очень низкой, если пропущено много данных. Как и любой другой метод оптимизации, данный алгоритм «локален», то есть процесс оптимизации сходится к локальному минимуму.

7.1.2. Методика заполнения неизвестных значений с учетом специфики предметной области

Как показывает практика, заполнение пропусков (неизвестных значений) в обучающей выборке возможно осуществлять с использованием некоторых приближений, полученных с учетом специфики предметной области, в нашем случае — неорганического материаловедения. Из периодического закона Д. И. Менделеева известно, что свойства химических элементов находятся в периодической зависимости от атомного номера.

Для решения проблемы заполнения неизвестных значений нами предложен следующий алгоритм с использованием метода «ближайших соседей» (рис. 7.1.1). Особо стоит отметить, что выполнение всех шагов согласно предлагаемому подходу, должно выполняться специалистом-химиком, обеспечивающим корректность проводимых интерполяций с учетом специфики предметной области. Во-первых, из обучающей выборки

удаляются признаки, имеющие больше 20 % (задается экспертом) пропусков, так как их информативность небольшая, но в случае заполнения пропусков они могут помешать правильному обучению системы. Заполнение же оставшихся пропусков предлагается реализовать по следующему алгоритму с учетом особенностей предметной области и периодического закона [299].

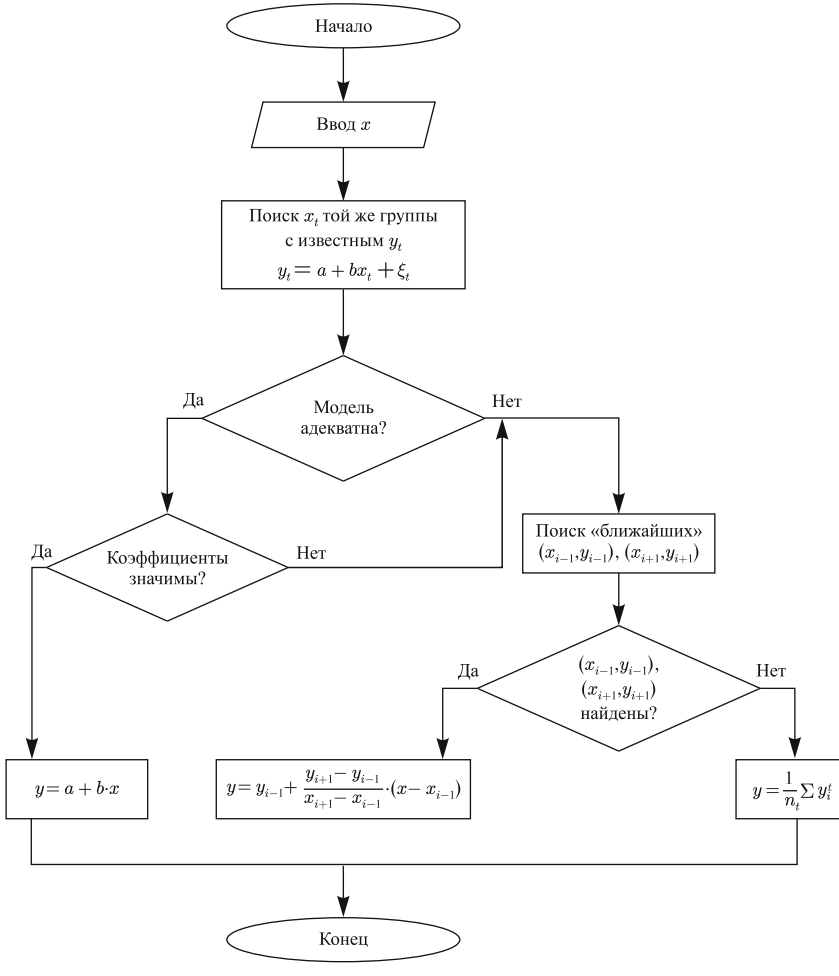


Рис. 7.1.1. Заполнение неизвестных значений с учетом специфики предметной области

Пусть неизвестно свойство y химического элемента x, тогда проводится поиск химических элементов, из той же группы периодической системы, у которых искомое свойство известно. Указанные химические элементы, согласно периодическому закону должны быть «близкими» по набору значений свойств к тому, у которого требуется заполнить пропуск. Т. е. получаем парную или простейшую линейную регрессию:

$$y_t = a + bx_t + \varepsilon_t ,$$

где x_t — вектор атомных номеров, y_t — вектор значений свойств.

Далее полученная линейная регрессия решается, например, методом наименьших квадратов (МНК) [303]. После проверки адекватности регрессионной модели, например, по критерию Фишера, и значимости коэффициентов, например, по критерию Стьюдента, принимается решение об использовании значения свойства, вычисленного с использованием найденной модели.

Если модель получается неадекватной (в силу возможных «выбросов» значений свойств внутри группы), то предлагается следующая схема вычисления недостающего значения. Находятся два «ближайших» элемента из той же группы с учетом их атомного номера, при этом относительное «расстояние» по атомным номерам между элементами не должно превышать заданного экспертом значения. Для найденных элементов (x_{i-1}, y_{i-1}) и (x_{i+1}, y_{i+1}) проводится линейная интерполяция, т. е. в результате неизвестное значение y вычисляется по формуле:

$$y = y_{i-1} + \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}} \cdot (x - x_{i-1}) .$$

Если подходящие элементы для линейной интерполяции не найдены, то неизвестное значение свойства элемента заменяется на среднее арифметическое значение этого признака у объектов с равным классообразующим признаком:

$$y = \frac{1}{n_t} \cdot \sum_{i=1}^{n_t} y_i^t ,$$

где t — класс объектов, к которому относится объект с пропуском в признаковом описании, n_t — количество объектов класса t в обучающей выборке, y_i^t — значение искомого свойства у i-го объекта того же класса. Или же этот признак исключается из обучающей выборки (выбор эксперта).

Таблица 7.1. Часть таблицы прогноза образования соединений состава $A^{I}B^{III}X_2$

Соединение (свойство)	*F*	$F/F_{табл}$
<Ln>I_3 ($T_{пл}$, K)	9,72	**3,95**
<Ln>I_3 ($S^o_{298\ K}$, кал/(моль·град)	0,94	0,38
<Ln>I_3 ($-\Delta H^o_{f\,298}$, ккал/моль)	122,94	**49,92**
<Ln>I_3 ($c^o_{p298.15}$, кал/моль·град)	0,00	0,00
<Ln>I_3 ($-\Delta G^o_{f\,298\ K}$, ккал/моль)	131,76	**53,50**
<Ln>Br_3 ($T_{пл}$, K)	33,52	**13,61**
<Ln>Br_3 ($S^o_{298\ K}$, кал/(моль·град)	7,14	**2,90**
<Ln>Br_3 ($-\Delta H^o_{f\,298}$, ккал/моль)	29,93	**12,15**
<Ln>Br_3 ($c^o_{p298.15}$, кал/моль·град)	1,18	0,48
<Ln>Br_3 ($-\Delta G^o_{f\,298\ K}$, ккал/моль)	21,12	**8,58**
<Ln>Cl_3 ($T_{пл}$, K)	0,27	0,11
<Ln>Cl_3 ($S^o_{298\ K}$, кал/(моль·град)	0,47	0,19
<Ln>Cl_3 ($-\Delta H^o_{f\,298}$, ккал/моль)	26,60	**10,80**
<Ln>Cl_3 ($c^o_{p298.15}$, кал/моль·град)	0,07	0,03
<Ln>Cl_3 ($-\Delta G^o_{f\,298\ K}$, ккал/моль)	20,23	**8,22**
<Ln>$_2O_3$ ($T_{пл}$, K)	8,24	**3,35**
<Ln>$_2O_3$ ($S^o_{298\ K}$, кал/(моль·град)	1,78	0,72
<Ln>$_2O_3$ ($-\Delta H^o_{f\,298}$, ккал/моль)	5,34	**2,17**
<Ln>$_2O_3$ ($c^o_{p298.15}$, кал/моль·град)	0,72	0,29
<Ln>$_2O_3$ ($-\Delta G^o_{f\,298\ K}$, ккал/моль)	72,39	**29,39**
<Ln>F_3 ($T_{пл}$, K)	29,86	**12,12**
<Ln>F_3 ($S^o_{298\ K}$, кал/(моль·град)	0,01	0,00
<Ln>F_3 ($-\Delta H^o_{f\,298}$, ккал/моль)	6,95	**2,82**
<Ln>F_3 ($c^o_{p298.15}$, кал/моль·град)	1,49	0,60
<Ln>F_3 ($-\Delta G^o_{f\,298\ K}$, ккал/моль)	0,46	0,19

Предложенная методика заполнения пропусков в свойствах химических элементов является комбинированным способом вычисления, основанным, прежде всего, на методе Бака, применяемом с учетом предметной области. По сравнению с методом, используемым в информационно-аналитической системе [300], основанным на методе вычисления безусловных средних, обеспечивается меньшее занижение дисперсии и увеличение корреляции, что дает в результате более качественную информацию для анализа алгоритмами распознавания образов, используемыми при компьютерном конструировании неорганических соединений.

Отметим, что строго говоря, предложенная методика может использоваться и для обработки отсутствующих значений в свойствах веществ. Например, для всех оксидов лантаноидов также наблюдается периодическая зависимость значений свойств от атомного номера химического элемента от лантана (La) до лютеция (Lu). В рамках исследований были изучены линейные регрессионные модели для различных свойств соединений лантаноидов с йодом, бромом, хлором, кислородом и фтором. При этом рассматривались следующие свойства:

- $T_{пл}$, K — температура плавления (в градусах по Кельвину);
- $S^o_{298\ K}$, кал/(моль·град) — энтропия при 298 K (кал/(моль·град));
- $\Delta H^o_{f\,298}$ — энтальпия образования при 298 K (ккал/моль);
- $c^o_{p298.15}$, кал/моль·град — теплоемкость при постоянном давлении и при 298 K (кал/моль·град);
- $-\Delta G^o_{f\,298\ K}$, ккал/моль — изобарный потенциал образования при 298 K (ккал/моль).

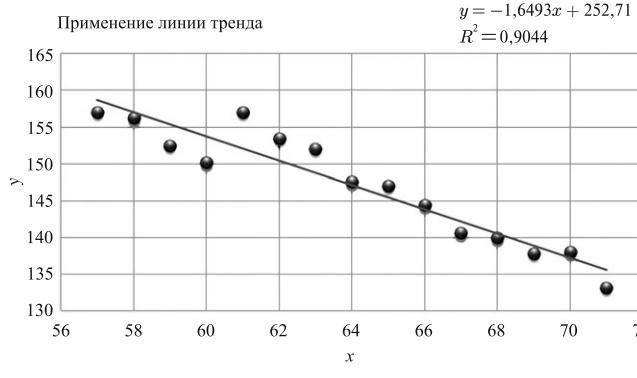
Для линейной регрессионной модели табличное значение критерия Фишера рассчитывалось с помощью встроенной в Excel функции FРАСПОБР(0,05;15;14), значение которой составило 2,46. Рассчитанное значение критерия Фишера для каждой модели сведено в табл. 7.1 (столбец *F*). Также приводится отношение *F/Fтабл*. Соответственно, если данное соотношение больше единицы, то модель адекватна по критерию Фишера.

Как видно из результатов сравнения *F* и *Fтабл* адекватно 14 из 25 регрессионных моделей. Самые высокие значения отношения *F/Fтабл* были обнаружены у следующих моделей:

- <Ln>I_3 ($-\Delta G^o_{f\,298\ K}$, ккал/моль) 131,76
- <Ln>I_3 ($-\Delta H^o_{f\,298}$, ккал/моль) 122,94
- <Ln>$_2O_3$ ($-\Delta G^o_{f\,298\ K}$, ккал/моль) 72,39

Приведем в качестве примера расчетов результаты построения линейной регрессионной модели для <Ln>I_3 ($-\Delta H^o_{f\,298}$, ккал/моль), см. рис. 7.1.2 и для <Ln>I_3 ($-\Delta G^o_{f\,298\ K}$, ккал/моль).

	X	*Y*(– Δ*Ho*, ккал/моль)	$(Y - Y_{сред})^2$	$Y_{расч}$	$(Y - Y_{расч})^2$
LaI_3	57	157	96,95684444	158,6972	2,880388
CeI_3	58	156,3	83,66151111	157,048	0,559534
PrI_3	59	152,5	28,58684444	155,3989	8,403443
NdI_3	60	150,2	9,282177778	153,7497	12,6005
PmI_3	61	157	96,95684444	152,1006	24,00444
SmI_3	62	153,4	39,02084444	150,4514	8,694141
EuI_3	63	152	23,49017778	148,8023	10,2255
GdI_3	64	147,6	0,199511111	147,1531	0,199705
TbI_3	65	147	0,023511111	145,504	2,23812
DyI_3	66	144,5	7,040177778	143,8548	0,416264
HoI_3	67	140,7	41,64551111	142,2057	2,267024
ErI_3	68	140	51,17017778	140,5565	0,309707
TmI_3	69	137,8	87,48484444	138,9074	1,226252
YbI_3	70	138,1	81,96284444	137,2582	0,708607
LuI_3	71	133,2	194,6955111	135,6091	5,803575
	$Y_{сред}$ =	**147,1533333**		**S =**	**80,5372**
				***a*1 =**	**–1,64915**
				***a*0 =**	**252,6988**

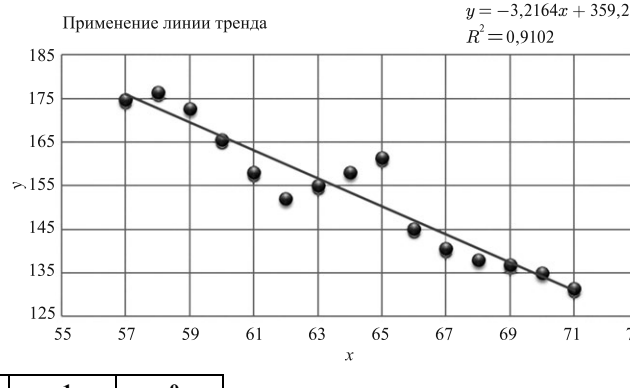


	***a*1**	***a*0**
	–1,64929	**252,7076**
	0,148747	9,541461
***r*^2**	**0,90437**	2,48901
F	**122,941**	13
	761,6401	80,53719
***Fтабл* =**	2,463003	

Модель адекватна *F*> $F_{табл}$

Рис. 7.1.2. Линейная регрессионная модель для <Ln>I3 ($-\Delta H_{of,298}$, ккал/моль)

	X	*Y*(–Δ*Go*, ккал/моль)	$(Y - Yc_{ред})^2$	$Y_{расч}$	$(Y - Y_{расч})^2$
LaI_3	57	174,665	452,7277508	175,9008248	1,527263049
CeI_3	58	176,489	533,674682	172,6846089	14,47339152
PrI_3	59	172,709	373,316498	169,468393	10,50153382
NdI_3	60	165,525	147,3164788	166,2521771	0,528786472
PmI_3	61	158	21,27423376	163,0359611	25,36090448
SmI_3	62	152,008	1,90329616	159,8197452	61,02336304
EuI_3	63	155	2,59983376	156,6035293	2,571306116
GdI_3	64	158,13	22,49035776	153,3873133	22,49307676
TbI_3	65	161,368	63,68678416	150,1710974	125,3706276
DyI_3	66	145,076	69,08269456	146,9548815	3,530195617
HoI_3	67	140,535	165,1893268	143,7386656	10,26347296
ErI_3	68	137,906	239,6799386	140,5224496	6,845808621
TmI_3	69	136,95	270,1946938	137,3062337	0,126902443
YbI_3	70	135	338,1038338	134,0900178	0,828067673
LuI_3	71	131,453	481,1266772	130,8738018	0,335470517
	***Yсред* =**	**153,3876**		***S* =**	**285,7801707**
				***a*1 =**	**–3,216215929**
				***a*0 =**	**359,2251328**



	***a*1**	***a*0**
	–3,21636	**359,2345**
	0,280198	17,9735
***r*^2**	**0,910199**	4,688613
F	**131,7643**	13
	2896,587	285,7802

***Fтабл* =** 2,463003

Модель адекватна *F*> *Fтабл*

Рис. 7.1.3. Линейная регрессионная модель для <Ln>I_3 ($-\Delta G^o_{f,\,298\,K}$, ккал/моль)

7.2. Этапы компьютерного конструирования новых соединений

Задача конструирования новых неорганических соединений сформулирована следующим образом [1, 153]: необходимо найти совокупность химических элементов и их соотношение (т. е. качественный и количественный состав) для создания (при заданных внешних условиях) определенной пространственной молекулярной или кристаллической структуры соединения, позволяющей реализовать необходимые функциональные свойства (формальная постановка задачи приведена в главе 1). Исходной информацией для расчетов должны быть только свойства химических элементов и данные о других уже изученных соединениях. Таким образом, речь идет о поиске зависимостей между свойствами систем (например, свойствами соединений) и свойствами химических элементов, образующих эти системы.

Одним из наиболее эффективных путей решения задачи конструирования многокомпонентных неорганических соединений является компьютерный анализ информации БД с целью поиска сложных закономерностей образования соединений определенных типов [1, 153] с использованием методов обучения ЭВМ распознаванию образов [186, 189, 192]. Найденные закономерности могут быть представлены в виде ассоциативной структуры данных, например, искусственной нейронной сети [192] или растущей пирамидальной сети [186], а также в форме булевского выражения, продукционных правил [186, 192], системы алгебраических уравнений [189, 192] и т. д. Переменными найденных закономерностей, как правило, являются свойства химических элементов. В качестве целевого параметра могут быть выбраны возможность образования соединения или тип его кристаллической структуры при заданных условиях, некоторое пороговое значение физического параметра, например, критическая температура перехода в сверхпроводящее состояние: выше 4,2 K или ниже и т. д.

Рассмотрим использование предложенного подхода для решения задач конструирования неорганических соединений, перспективных для поиска новых веществ для электроники. В этом случае речь идет о нахождении (среди возможных комбинаций различных элементов) аналогов уже известных соединений, обладающих искомыми свойствами [190].

Первый этап компьютерного конструирования новых соединений — это экспертный анализ информации баз данных по свойствам веществ для электроники и выбор соединений-прототипов. Систематизированную информацию на этом этапе предоставляет описанная выше интегрированная система БД. Учитывая важность химического состава и кристаллической

структуры для проявления физических свойств, было решено вести поиск аналогов соединений-прототипов именно по этим параметрам: составу и кристаллической структуре.

Следующий этап компьютерного конструирования — это отбор информации об известных аналогах по составу и/или типу кристаллической структуры в базе данных по свойствам неорганических соединений «Фазы». Например, для того, чтобы осуществить прогноз еще неисследованных селенидных систем, в которых при обычных условиях образуется соединение состава $ABSe_2$ (A и B — здесь и далее различные химические элементы), в БД «Фазы» запрашивается информация об известных системах с селеном, в которых образуются соединения прогнозируемого состава, и о системах, в которых при нормальных условиях такие селениды не обнаружены. Каждая система описывается в виде набора значений свойств химических элементов, входящих в ее состав. Данные о свойствах химических элементов извлекаются из БД «Elements». Как правило, используется множество самых различных свойств элементов и/или их простых соединений (в данном случае простых селенидов). Результатом этого этапа является матрица, строками которой являются описания систем в терминах свойств элементов и/или их простых соединений и указание об их принадлежности к тому или иному классу систем (в данном случае — к классам систем с образованием и без образования соединений состава $ABSe_2$).

Далее с помощью программ обучения ЭВМ [186] или [192] проводится анализ полученной матрицы и выделение интервалов изменения значений свойств, которые соответствуют различным классам систем. Результатом этого этапа является закономерность, разделяющая системы на разные классы.

На заключительном этапе в найденную закономерность подставляются наборы значений свойств элементов — компонентов еще неисследованных систем, и исследователь получает прогноз, будет ли образовываться в данной системе соединение заданного состава или нет.

Точно так же можно получить прогноз соединений с определенным типом кристаллической структуры или с параметрами, значения которых находятся в определенном интервале.

На основе системного анализа процесса компьютерного конструирования неорганических соединений разработана методика использования интегрированной ИС СНВМ в качестве источника данных информационно-аналитической системы, (ИАС) используемой для поддержки принятия решений при исследовании неорганических соединений (рис. 7.2.1) [193].

Прогнозирующая ИС используется, если в интегрированной ИС нет данных по соединениям с нужными свойствами. После экспериментальной проверки результатов прогноза, информация помещается в разрабо-

танную интегрированную ИС, пополняя соответствующие БД. При этом если эти данные не совпадают с результатом прогноза, то пользователь может инициировать переобучение системы с учетом новых данных. За счет использования бо́льшей обучающей системе, вероятно, удастся построить лучшую закономерность и тем самым повысить точность будущих прогнозов.

Приведем некоторые примеры компьютерного конструирования новых соединений, перспективных для использования в электронике, согласно предложенной схеме использования интегрированной ИС в СППР.

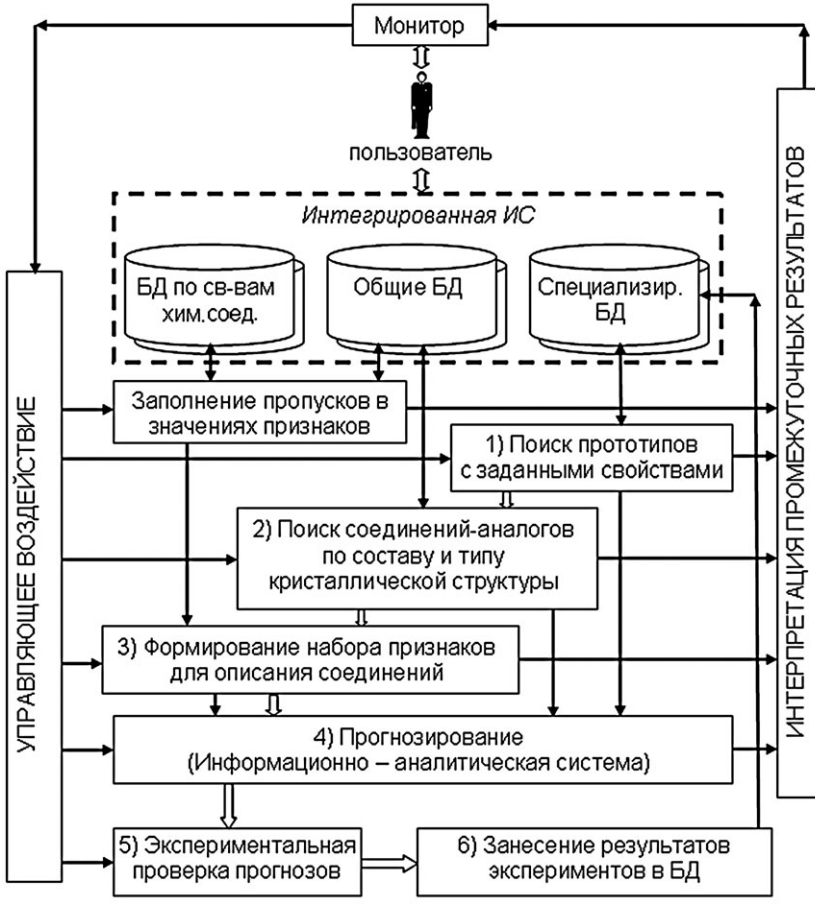


Рис. 7.2.1. Методика использования интегрированной ИС СНВМ в ИАС

7.3. Перспективные полупроводники ABX_2

Еще полвека назад академик А. Ф. Иоффе отметил в своей монографии [196], что свойства полупроводников, в первую очередь определяются их химической природой, и поэтому могут быть предсказаны с этих позиций. Химическая природа полупроводниковых фаз непосредственно связана с их составом и кристаллической структурой.

На основе анализа информации БД «Bandgap» и «Диаграмма» были отобраны тройные полупроводниковые соединения-прототипы с различными стехиометрическими составами и осуществлено конструирование еще неполученных фаз-аналогов полупроводниковых соединений. Матрица информации для компьютерного анализа формировалась на основе данных БД «Фазы».

Программная система [186] была использована для прогноза новых соединений состава ABX_2 (X = S, Se, Te), перспективных для поиска новых полупроводниковых и нелинейно-оптических веществ. Для описания химических соединений были использованы параметры химических элементов, хранящиеся в БД «Elements»: распределение электронов по энергетическим оболочкам изолированных атомов, электроотрицательности по Полингу, первые три потенциала ионизации, ковалентные радиусы и т. д., а также энтальпии образования и энтропии соответствующих простых халькогенидов (сульфидов или селенидов) при стандартных условиях.

В табл. 7.2 даны результаты сравнения полученных прогнозов с экспериментальными данными. Приняты следующие обозначения: «+» — прогноз образования соединений состава ABX_2 при нормальных условиях; «–» — прогноз отсутствия соединений состава ABX_2 в системе A–B–X при нормальных условиях; «⊕» — соединение состава ABX_2 существует и этот факт использован для обучения ЭВМ; «↔» — соединение состава ABX_2 не образуется в системе A–B–X и этот факт использован для обучения ЭВМ; «©» — прогноз образования соединения состава ABX_2 подтвержден экспериментом; «О» — прогноз отсутствия соединения состава ABX_2 в системе A–B–X подтвержден экспериментом; «∅» — прогноз отсутствия соединения состава ABX_2 в системе A–B–X не подтвержден экспериментом; пустые клетки — неопределенный прогноз. Из 61 проверенного прогноза только три оказались неправильными.

Далее был осуществлен прогноз типа кристаллической структуры при нормальном давлении и комнатной температуре для соединений, предсказанных выше [1, 155]. Для соединений состава ABX_2 с кристаллической структурой халькопирита была решена задача прогноза ширины запрещенной зоны с использованием программных комплексов обучения ЭВМ [186, 192]. Информация для компьютерного анализа о ширине запрещенной зоны известных халькопиритов этого состава была получена из БД «Bandgap».

Таблица 7.2. Часть таблицы прогноза образования соединений состава $A^{I}B^{III}X_2$

X	S					Se					Te							
B^{III} \ A^{I}	Na	Rb	Ag	Cs	Tl	Li	Na	K	Rb	Ag	Li	Na	K	Cu	Rb	Ag	Cs	Tl
B	⊕	©	+	©	⊕	–	–	+	+	+	–	–	+	+	+	+	+	+
Al	⊕		⊕	+	⊕		⊕	⊕	+	⊕	©	⊕	⊕	⊕	+	⊕	+	©
Sc	⊕	–	–	–	⊕	+	+	+	+	⊕	+	+	+	+	+	+	+	+
Ti	©	⊕	+	⊕	+	⊕	©	+	+	+	+	+	+	+	+	+	+	+
V	⊕	–	–	–	+	©	⊕	+	+	+	©	©	+	–	+	+	+	+
Cr	⊕	⊕	⊕	?	⊕	+	⊕	+	⊕	⊕	+	+	+		–	©	+	⊕
Mn	+	–	–	–	+	©	©	+	©	+	©	©	©	–	©	©	©	+
Fe	+	⊕	⊕	⊕	⊕	+	+	⊕	⊕	⊕	+	+	+	⊕	+	⊕	+	©
Co	+				+	+	+	+	+	+	+	+	+	–	+	©	+	+
Ni	+	–	–	–	+	+	⊕	+	+	⊕	+	+	+	–	+	⊕	+	+
Ga	⊕	⊕	⊕	©	©		+	⊕	+	⊕	+	⊕	⊕	⊕		⊕	+	⊕
As	⊕		⊕		⊕	⊕	⊕	⊕	⊕	⊕	+	+	+	↔		⊕	+	⊕
Y	⊕		⊕		⊕	⊕	⊕	+	+	⊕	+	+	+	⊕	+	⊕	+	⊕
Rh	+				+	+	+	+	+	+	+	+	?	–	+	–	+	+
In	⊕	©	⊕	©	⊕	⊕	⊕	⊕	©	⊕	⊕	⊕	⊕	⊕	©	⊕	+	©
Sb	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	©	O	O	⊕	⊕	⊕	⊕
La	⊕	⊕	⊕	⊕	↔	⊕	⊕	+	©	∅	+	+	+	⊕	+	+	+	↔
Ce	⊕	⊕	⊕	⊕	↔	⊕	⊕	+	©	↔	+	+	+	⊕	+	+	+	+
Pr	⊕	⊕	⊕	↔	↔	+	⊕	+	©	↔	+	+	+	⊕	+	+	+	⊕
Nd	⊕	⊕	⊕	↔	↔	©	⊕	+	©	⊕	+	+	+	⊕	+	+	+	⊕
Pm	+					+	+	+	+	+	+	+	+	+	+	+	+	+
Sm	⊕	⊕	⊕		⊕	©	⊕	+	©	↔	+	+	©	⊕	+	+	+	⊕
Eu	⊕	⊕	⊕	–	⊕	+	⊕	+	+	↔	+	+	+	+	+	+	+	↔
Gd	⊕	⊕	⊕	+	⊕	⊕	⊕	+	©	∅	+	+	+	⊕	+	⊕	+	⊕
Tb	⊕	⊕	⊕	+	⊕	⊕	⊕	+	©	∅	+	+	+	⊕	+	+	+	⊕
Dy	⊕	↔	⊕		⊕	⊕	⊕	+	+	⊕	+	+	+	⊕	+	⊕	+	⊕
Ho	⊕	©	⊕		⊕	⊕	⊕	+	©	⊕	+	+	+	⊕	+	⊕	+	⊕
Er	⊕	©	⊕		⊕	⊕	⊕	+	©	⊕	+	+	©	⊕	+	⊕	+	⊕
Tm	⊕	©	⊕		⊕	+	+	+	+	⊕	+	+	+	⊕	+	⊕	+	⊕
Yb	⊕	©	⊕	+	⊕	+	©	©	+	⊕	+	+	+	⊕	+	+	+	⊕
Lu	⊕	©	©	+	⊕	+	+	+	©	⊕	+	+	+	⊕	+	⊕	+	⊕
Tl	+	©	⊕	©	+	©	©	©	+	⊕	+	+	+	⊕	+	⊕	+	+
Bi	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕	⊕
Ac	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+

Для описания соединений были взяты следующие функции от параметров химических элементов:

- отношение валентности к ковалентному радиусу;
- средняя атомная масса m_{cp}, вычисляемая по формуле

$$m_{cp} = (m_A + m_B + 2m_X)/4;$$

- среднее отношение первого потенциала ионизации к валентности I/z, вычисляемое по формуле:

$$I/z = \{(I_z/z)_A + (I_z/z)_B + (I_z/z)_X\}/4.$$

В табл. 7.3 даны результаты сравнения прогнозируемых значений ширины запрещенной зоны с экспериментальными значениями для различных алгоритмов, программы которых включены в системы [186, 192].

Приняты следующие обозначения:

- класс соединений:

  «**1**» — $E_g \le 2$ эВ;
  «**2**» — $E_g > 2$ эВ;

- алгоритмы:

  «**I**» — алгоритмы вычисления оценок;
  «**II**» — линейный дискриминант Фишера;
  «**III**» — линейная машина;
  «**IV**» — логические закономерности;
  «**V**» — многослойный персептрон;
  «**VI**» — Q ближайших соседей;
  «**VII**» — метод опорных векторов;
  «**VIII**» — статистически взвешенные синдромы;
  «**IX**» — голосование по тупиковым тестам [183];
  «**X**» — Confor [186].

В табл. 7.4 приведена точность прогноза ширины запрещенной зоны известных халькопиритов с использованием вышеуказанных программ обучения ЭВМ [186, 192]. Экзаменационное распознавание проводилось на изначально неиспользованной для обучения части массива обучающей выборки, после чего проводилось переобучение по всей обучающей выборке. Следует отметить, что точность прогноза достаточно высокая — выше 77 %, что свидетельствует о возможности правильной оценки такой важного для характеристики полупроводниковых веществ параметра как ширина запрещенной зоны.

На основе анализа результатов табл. 7.3 сделан вывод, что для повышения точности прогноза следует принимать решение о принадлежности соединения к тому или иному классу на основе «голосования» результатов прогноза с использованием различных алгоритмов [192]. В случае решаемой задачи это позволило получить 100 %-ю точность прогноза.

Таблица 7.3. Результаты прогноза ширины запрещенной зоны халькопиритов состава ABX_2

Состав	Класс (эксп).	Прогноз									
		I	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**
$CuAlS_2$	**2**	2	2	2	2	2	2	2	2	2	2
$CuGaS_2$	**2**	2	2	2	2	2	2	2	2	2	2
$CuInS_2$	**1**	1	1	1	1	2	1	1	1	2	1
$CuAlSe_2$	**2**	1	1	2	2	2	2	2	2	2	2
$CuGaSe_2$	**1**	1	1	1	1	2	1	1	1	1	1
$CuInSe_2$	**1**	1	1	1	1	2	1	1	1	1	1
$CuAlTe_2$	**1**	1	1	1	1	1	1	1	1	1	1
$CuGaTe_2$	**1**	1	1	1	1	1	1	1	1	1	1
$CuInTe_2$	**1**	1	1	1	1	1	1	1	1	1	1
$AgGaS_2$	**1**	1	1	1	1	2	1	1	1	2	1
$AgGaSe_2$	**1**	1	1	1	1	2	1	1	1	1	1
$AgInSe_2$	**1**	1	1	1	1	2	1	1	1	1	1
$AgAlTe_2$	**1**	1	1	1	1	1	1	1	1	1	1
$AgGaTe_2$	**1**	1	1	1	1	1	1	1	1	1	1
$AgInTe_2$	**1**	1	1	1	1	1	1	1	1	1	1
$ZnSiP_2$	**2**	2	2	2	2	2	2	2	2	2	2
$ZnSiAs_2$	**2**	1	1	2	2	2	2	2	2	2	2
$ZnGeP_2$	**1**	1	1	1	?	1	1	1	2	2	1
$ZnGeAs_2$	**1**	1	1	1	1	1	1	1	1	1	1
$ZnSnP_2$	**1**	1	1	1	1	1	1	1	1	1	1
$ZnSnAs_2$	**1**	1	1	1	1	1	1	1	1	1	1
$CdSiP_2$	**2**	2	2	2	2	2	2	2	2	2	2
$CdGeP_2$	**1**	1	1	1	1	1	1	1	1	1	1
$CdGeAs_2$	**1**	1	1	1	1	1	1	1	1	1	1
$CdSnP_2$	**1**	1	1	1	1	1	1	1	1	1	1
$CdSnAs_2$	**1**	1	1	1	1	1	1	1	1	1	1
$AgInS_2$	**1**	1	1	1	1	2	1	1	1	1	1
$CdSiAs_2$	**1**	1	1	1	1	1	1	1	1	1	1

Таблица 7.4. Точность прогноза ширины запрещенной зоны халькопиритов различными методами

Алгоритм	Точность, %
алгоритмы вычисления оценок	87.1
линейный дискриминант Фишера	83.9
линейная машина	96.8
логические закономерности	100
многослойный персептрон	87.1
Q ближайших соседей	77.4
метод опорных векторов	100
статистически взвешенные синдромы	80.6
голосование по тупиковым тестам	77.4
Confor	100

7.4. Перспективные диэлектрики $A_2B_2(XO_4)_3$

Известно, что линейный электрооптический эффект, пьезоэлектрический эффект и эффект генерации второй гармоники возможны только в нецентросимметричных (ацентричных) кристаллах. Рез и Поплавко [210], отмечая необходимость поиска нецентросимметричных диэлектрических кристаллов для электрооптических, нелинейнооптических и пьезоэлектрических приложений, указывали на важность пространственной группы в проявлении этих свойств.

Для поиска кристаллоструктурных аналогов веществ этих классов использовалась информация БД по свойствам акустооптических, электрооптических и нелинейнооптических веществ «Кристалл», входящая в состав разработанной интегрированной информационной системы. Примеры для обучения ЭВМ были взяты из БД по свойствам неорганических соединений «Фазы», а информация о свойствах химических элементов для описания соединений — из БД «Elements». Для поиска закономерностей в информации и прогноза использовалась система программ [186].

Были выбраны соединения состава $A_2B_2(XO_4)_3$ с кристаллической структурой лангбейнита (пр. гр. $P2_13$), которые принадлежат к одному из перспективных классов пьезоэлектрических, сегнетоэлектрических, нелинейнооптических, электрооптических и люминесцентных веществ [210, 211, 212, 213]. Табл. 7.5 содержит результаты экспериментальной проверки

полученных прогнозов новых лангбейнитов. Приняты следующие обозначения: «**L**» — прогноз образования соединения с кристаллической структурой лангбейнита; «**K**» — прогноз образования соединения со структурой типа $K_2Zn_2(MoO_4)_3$; «*****» — прогноз отсутствия соединения состава $A_2B_2(XO_4)_3$ в системе A–B–X–O при нормальных условиях получения; «–» — прогноз соединения с типом кристаллической структуры, отличным от лангбейнита или $K_2Zn_2(MoO_4)_3$**; «(L)»**, «**(K)**» — соединение с соответствующим типом кристаллической структуры при нормальных условиях было известно и информация о нем была использована для обучения ЭВМ; «**↔**» — соединение с типом кристаллической структуры, отличным от лангбейнита или $K_2Zn_2(MoO_4)_3$, получено при нормальных условиях и информация о нем была использована для обучения ЭВМ; «**(*)**» — соединение состава $A_2B_2(XO_4)_3$ не получено в системе A–B–X–O и информация об этом была использована для обучения ЭВМ; пустые ячейки и «**?**» — неопределенный результат; значок «**©**» — прогноз подтвержден экспериментом; значок «**¢**» — прогноз не подтвержден экспериментом. Даже для таких сложных по составу соединений только 5 из 17 проверенных результатов не совпали с экспериментом.

Таблица 7.5. Часть таблицы прогноза типа кристаллической структуры соединений состава $A_2B_2(XO_4)_3$

X	**S**					**Cr**					**Mo**					**W**				
A \ **B**	**Na**	**K**	**Rb**	**Cs**	**Tl**	**Na**	**K**	**Rb**	**Cs**	**Tl**	**Na**	**K**	**Rb**	**Cs**	**Tl**	**Na**	**K**	**Rb**	**Cs**	**Tl**
Mg	L¢	(L)	(L)	(*)	L	L	L©		L©	L©	K¢	(K)	(L)	(L)	(L)		↔	(L)	L©	L
Ca	(*)	(L)	L©	(L)	(*)			L	L	L	(*)	?	?	?	?	(*)	*	?	?	?
Mn	(*)	(L)	(L)	L	(L)	L		(L)	L©	L	K	↔	(L)	(L)	(L)					
Fe	*	L©	L©		(L)	L	K	L	L	L	K	K	?	?	?		K			
Co	(*)	(L)	L©		(L)	L	K	L	L	L		(K)	(L)	(L)	↔		K			
Ni	(*)	(L)	L©	L	L	L		L	L	L	K	(K)	(L)	(L)	(L)					
Cu	(*)		L	*	L	L	K	L	L	L	K	(K)	?	?	?		K			
Zn	*	(L)	L	*	L	L	K	L	L	L	↔	(K)	(K)	–	(K)		K¢			
Sr	(*)	?	?		(*)	*	*	?	?	?	(*)	?	?	?	?	(*)	*	*	*	*
Cd	(*)	(L)	(L)		(L)							K	↔	(L)	K¢	(*)		L	L	L
Ba	(*)		(*)	(*)	(*)	*	*				(*)		*	*	*	*	*			
Pb					*	(*)	*	*©	*	*	(*)	*©	(*)	*¢	*	(*)	*	(*)	(*)	*

7.5. Прогноз образования сегнетоэлектрических хлоридов A_2BCl_4

Согласно информации ИС «Кристалл» соединения состава A_2BCl_4 относятся к группе сегнетоэлектрических кристаллов типа K_2SO_4, охватывающей около десятка соединений, наиболее известными из которых являются Rb_2ZnCl_4, K_2ZnCl_4, Rb_2ZnBr_4, $(NH_4)_2ZnCl_4$. Для этих соединений характерно то, что переход от сегнетоэлектрической фазы к параэлектрической фазе происходит через промежуточную несоразмерную фазу [304]. У кристаллов A_2BCl_4 выявлены электрооптические и нелинейные оптические свойства [305]. Например, у Rb_2ZnCl_4 значения электрооптических коэффициентов составляют при 186.6 К для $\lambda = 0.633$ мкм: $r_{11} = 1.0 \cdot 10^{-12}$ м/В, $r_{23} = 1.5 \cdot 10^{-12}$ м/В, $r_{33} = 0.3 \cdot 10^{-12}$ м/В. Значения нелинейных оптических коэффициентов равны при температуре 100 К для $\lambda = 1.064$ мкм: $d_{33} = 0.085 \cdot 10^{-12}$ м/В, $d_{22} = 0.075 \cdot 10^{-12}$ м/В, $d_{24} = 0.069 \cdot 10^{-12}$ м/В, а величины компонент тензора Миллера при температуре 190 К для $\lambda = 1.064$ мкм: $\delta_{33} = 0.18$ м^2/Кл, $\delta_{32} = 0.16$ м^2/Кл, $\delta_{24} = 0.14$ м^2/Кл. По известным данным кристаллы Rb_2ZnCl_4 генерируют вторую оптическую гармонику [305]. С целью поиска еще не синтезированных сегнетоэлектрических хлоридов нами проведено прогнозирование новых соединений состава A_2BCl_4 (A и B — разные металлы).

При отборе примеров соединений для компьютерного анализа использовались ИС СНВМ, разработанные в ИМЕТ РАН [268]. На основе анализа информации ИС «Фазы» была сформирована выборка, содержащая 68 примеров соединений A_2BCl_4 и 29 примеров систем без образования соединений состава A_2BCl_4.

Стоит отметить, что отбор свойств элементов для включения в обучающую выборку (наиболее сложная и влияющая на качество прогнозирования задача) базировался на основе физико-химических представлений специалистов-материаловедов о природе изучаемых фаз. Так в исходный набор свойств были включены 67 параметров элементов A и B и простых хлоридов составов ACl и BCl_2, часть из которых представлена в таблице (табл. 7.6). Для компьютерного анализа данных был использован комплекс алгоритмов распознавания образов по прецедентам, включенный в разработанную в ИМЕТ РАН информационно-аналитическую систему [308]. При решении задачи проводился отбор наиболее точных алгоритмов. Для этого применялось экзаменационное распознавание со скользящим контролем (cross validation) на материале обучающей выборки, которое является традиционным средством оценки качества обучения ЭВМ [192].

Методика использования скользящего контроля следующая. Обучающая выборка разбивается N раз различными способами на две непересе-

кающиеся подвыборки: обучающую подвыборку длины m, и контрольную подвыборку длины k. Для каждого i-го разбиения ($i = 1, \ldots, N$) строится алгоритм распознавания и вычисляется процент ошибок по формуле

$$Q_i = \frac{Q_{err\_i}}{Q_{общ\_i}} \cdot 100\%,$$

где Q_{err_i} — количество неверных прогнозов на i-м разбиении, $Q_{общ\_i}$ — общее количество прогнозов на i-м разбиении. Соответственно, среднее арифметическое значений Q_i по всем разбиениям называется оценкой скользящего контроля:

$$Q = \frac{1}{N}\sum_{i=1}^{N} Q_i$$

Для повышения точности прогнозирования соединений была использована стратегия коллективов алгоритмов [192]. Как правило, использование стратегии коллективов алгоритмов позволяет улучшить точность прогнозирования за счет взаимной компенсации недостатков одного алгоритма преимуществами других. Для оценки точности «коллективных» алгоритмов применялось экзаменационное распознавание 50 примеров, случайно

Таблица 7.6. Некоторые свойства для описания соединений в системе A–B–Cl

Свойство	Элемент A	Элемент B
E8 Химический потенциал Miedema	**1**	**32**
I8 Термическая проводимлсть	**2**	**33**
S6 Удаленный электрон ядра по Шуберту	**3**	**34**
S5 Удаленный валентный электрон по Шуберту	**4**	**35**
E2 Электроотрицательность по Полингу	**5**	**36**
E5 Энергия первичной ионизации	**6**	**37**
E6 Энергия вторичной ионизации	**7**	**38**
E7 Энергия третичной ионизации	**8**	**39**
C5 Энтальпия атомизации	**9**	**40**
I11 Энтропия твердого тела	**10**	**41**
G1 Номер группы	**11**	**42**
I10 Молярная теплоемкость	**23**	**54**
A5 Квантовое число	**24**	**55**
S11Ионные радиусы	**25**	**56**
S1 Псевдопотенциальные радиусы	**26**	**57**
C2 Температура кипения	**27**	**58**
C1 Температура плавления	**28**	**59**
G2 Число валентных электронов	**29**	**60**
Tm – температура плавления простого хлорида	**63**	
$S°_{298\,K}$ – энтропия простого хлорида при 298 K	**64**	
$-\Delta H°_{f298}$ – теплота образования простого хлорида при 298 K	**65**	
$c°_{p298.15}$ – изобарная теплоемкость простого хлорида при 298 K	**66**	

выбранных из обучающей выборки и не использованных в обучении ЭВМ (на завершающем этапе прогнозирования контрольные примеры возвращались в обучающую выборку).

Найденная классифицирующая закономерность была использована для прогнозирования еще не полученных соединений. Следует отметить, что для прогнозирования новых соединений использовались только значения свойств компонентов (химических элементов).

Таблица 7.7. Оценка достоверности прогнозирования возможности образования соединений A_2BCl_4 с использованием различных методов распознавания образов

Алгоритм	Достоверность экзаменационного распознавания со скользящим контролем, %	Примечания
Алгоритм вычисления оценок	**73.2**	Плохо распознаны объекты класса «отсутствие соединения»
Генетический метод	**82.3**	Плохо распознаны объекты класса «отсутствие соединения»
Двумерные линейные разделители	**67.0**	Плохо распознаны объекты класса «отсутствие соединения»
Метод бинарных решающих деревьев	**79.4**	
Линейный дискриминант Фишера	**75.3**	
Линейная машина	**77.3**	
Логические закономерности классов	**68.8**	
Логические закономерности	**80.2**	
Мультипликативная нейронная сеть	**74.0**	
Многослойный персептрон (Сигмоид)	**74.0**	
Нейронная сеть	**77.1**	
k ближайших соседей	**75.0**	
Метод опорных векторов	**83.3**	
Статистически взвешенные синдромы	**82.3**	
Голосование по тупиковым тестам	**76.0**	
ConFor	**57.0**	

Было установлено, что наиболее важными для классификации систем по признаку существования или отсутствия соединений являются: изобарный потенциал образования простого галогенида BCl_2, температура плавления элемента A и псевдопотенциальный радиус элемента B.

На основе анализа результатов экзаменационного распознавания (табл. 7.7) можно сделать вывод, что лучшие результаты прогнозирования могут быть получены с использованием программ на основе алгоритмов «метод бинарных решающих деревьев», «логические закономерности», «метод опорных векторов» и «статистически взвешенные синдромы». Именно эти алгоритмы были использованы при принятии коллективного решения.

С целью дальнейшего увеличения точности прогнозирования были проведены компьютерные эксперименты по поиску эффективных методов принятия коллективного решения с использованием наиболее точных алгоритмов распознавания образов по прецедентам. Лучшие результаты были получены при использовании алгоритма «выпуклый стабилизатор» (табл. 7.8).

Таблица 7.8. Оценка достоверности прогнозирования возможности образования соединений A_2BCl_4 с использованием различных методов принятия коллективного решения

Алгоритм	**Достоверность экзаменационного распознавания случайно выбранных 50 объектов, не включенных в обучающую выборку, %**	**Примечания**
Алгебраический корректор	78.0	
Метод Байеса	75.0	
Области компетенции	65.0	Плохо распознаны объекты класса «отсутствие соединения»
Шаблоны принятия решений	80.0	
Динамический метод Вудса	74.0	
Комплексный комитетный метод — голосование по большинству	75.0	
Комплексный комитетный метод — усреднение	80.0	
Выпуклый стабилизатор	85.0	
Обобщенный полиномиальный корректор	75.0	

В табл. 7.9 даны прогнозы новых соединений состава A_2BCl_4. Приняты следующие обозначения: 1 — прогноз возможности образования соединения состава A_2BCl_4 при обычных условиях; 2 — прогноз отсутствия соединения A_2BCl_4 в системе A–B–Cl. Значком «#» обозначены ранее изученные системы, информация о которых использована для обучения ЭВМ.

Таблица 7.9. Прогнозы возможности образования соединений состава A_2BCl_4

B \ A	Li	Na	K	Rb	Cs	Tl
Be	#1	#1	#1	#1	#1	#1
Mg	1	#1	#1	#1	#1	#2
Ca	#2	#2	#2	#2	#1	2
Ti	#1	#1	#1	#1	#1	1
V	1	1	#1	#1	#2	1
Cr	#1	#1	#1	#1	#1	1
Mn	1	#1	#1	#1	#1	#2
Fe	#1	1	#1	#1	#1	1
Co	#1	#1	#1	#1	#1	#1
Ni	1	#2	1	1	1	1
Cu	#2	#2	#1	#1	#1	1
Zn	1	#1	#1	#1	#1	#1
Sr	2	#2	#1	#1	#2	#2
Cd	1	#1	#1	#1	#1	#1
Sn	#2	#2	#2	#1	#1	#2
Ba	#2	#2	#1	#1	#1	#2
Eu	#2	#2	#1	#2	#2	#2
Yb	1	#2	#2	1	#1	2
Hg	#1	#1	#1	#1	#1	1
Pb	#2	#2	#1	#1	#1	#1

7.6. Прогноз образования соединений состава AB_2X_4

Большинство соединений состава AB_2X_4 относятся к полупроводниковым соединениям. Халькогенидные шпинели состава AB_2X_4 (X = S, Se, Te) представляют интерес для поиска новых магнитных полупроводников, подобных известным фазам состава $CdCr_2S_4$, $CdCr_2Se_4$, $HgCr_2Se_4$, $ZnCr_2Se_4$,

$CuCr_2Se_4$, $FeCr_2S_4$ и т. д., открытым в 1960-е годы. Халькогенидные магнитные полупроводники со структурой шпинели нашли применение в полупроводниковых приборах с управлением магнитным полем, например, в управляемых МДП-структурах, в приборах, использующих гигантское (до $5 \cdot 10^6$ град/см) фарадеевское вращение плоскости поляризации в магнитном поле, в квантовых приемниках и элементах памяти, работающих на принципе сильного фотомагнетизма в магнитном поле. Халькогенидные шпинели могут использоваться также в узкополостных источниках света, управляемых магнитным полем. Перспективно применение халькошпинелей в интегральных схемах, в которых один участок используется как активное полупроводниковое устройство, а другой — как магнитный микроволновой прибор, а также в устройствах, где существенна взаимосвязь электрических, магнитных и оптических свойств. Интерес с халькошпинелям вызывает и обнаружение слабой сверхпроводимости у некоторых из этих фаз: $CuRh_2S_4$ ($T_c = 4.8$ K), $CuRh_2Se_4$ ($T_c = 3.49$ K), CuV_2S_4 ($T_c = 4.45$ K). Халькошпинели рассматриваются как перспективные термоэлектрические материалы [306].

В ИС «Фазы» хранится информация о более тысячи соединений состава AB_2X_4. Для примерно 2/3 этих соединений существуют данные о кристаллической структуре. Неоднократно предпринимались попытки поиска критериев образования соединений этого состава, а также критериев, позволяющих разделить фазы с различными кристаллическими структурами.

В последние годы были синтезированы и изучены сотни новых халькогенидных соединений подобного состава, что позволило уточнить прогнозы возможности образования новых соединений этого состава и типа их кристаллической структуры при обычных условиях за счет использования новых данных.

Для обучения были использованы 835 примеров образования соединений (класс 1) и 154 отсутствия соединений состава AB_2X_4 (X = S, Se или Te) (класс 2) в системах $AX–B_2X_3$, $AX_2–BX$ и $A_2X–BX_3$ при обычных условиях. Информация была извлечена из ИС «Фазы». Большинство халькогенидных соединений состава AB_2X_4 (X = S, Se, Te) кристаллизуются в структурных типах Th_3P_4 (пр. гр. I4(-)3d, Z = 4), $CaFe_2O_4$ (пр. гр. Pnam, Z = 4), шпинели (пр. гр. Fd3m, Z = 8), $PbGa_2Se_4$ (пр. гр. Fddd, Z = 32), $MnEr_2S_4$ (пр. гр. $Cmc2_1$, Z = 4), $NiCr_2S_4$ (пр. гр. I2/m, Z = 2), $CaHo_2Se_4$ (пр. гр. R32, Z = 0.5), $CdAl_2S_4$ (пр. гр. I4(-), Z = 2), $PbBi_2S_4$ (пр. гр. $P2_12_12_1$, Z = 4), $GeSr_2S_4$ (пр. гр. $P2_1/m$, Z = 2), $PbCr_2S_4$ (пр. гр. P6, Z = 9), TlSe (пр. гр. I4/mcm, Z = 2), сфалерита (пр. гр. F4(-)3m, Z = 1) и т. д. Анализ информации из ИС «Фазы» показывает, что наиболее распространенными являются структурные типы Th_3P_4, шпинели и $CaFe_2O_4$. На рис. 7.6.1 показана гистограмма распространенности типов кристаллической структуры.

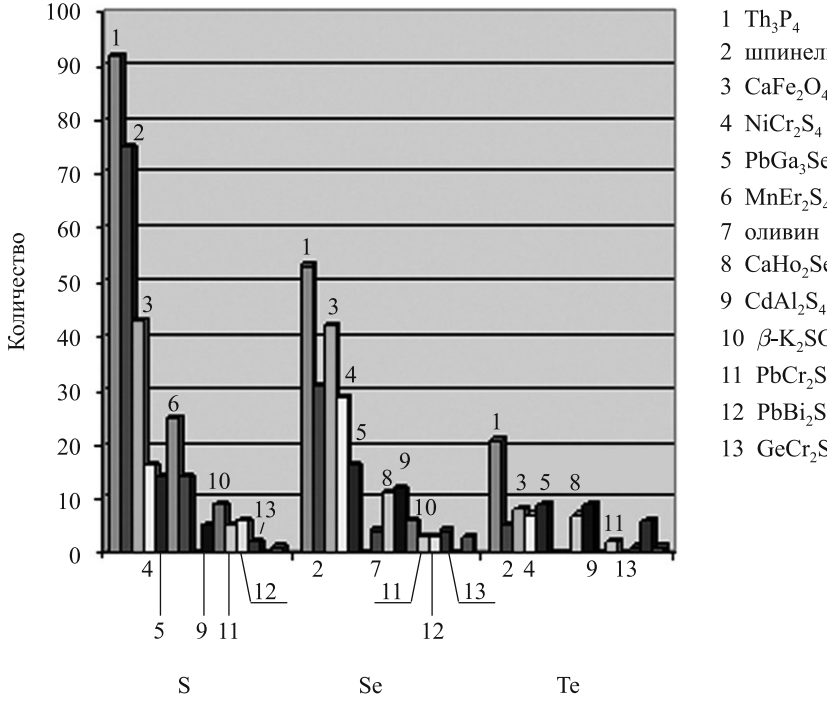


Рис. 7.6.1. Гистограмма распространенности типов кристаллической структуры AB_2X_4

Поиск гипотез образования различных кристаллических фаз проводился в многомерных пространствах свойств компонентов, перечень которых дан в табл. 7.10. Химические системы представлялись в виде набора значений свойств химических элементов A, B и X. Информация о свойствах химических элементов была взята из БД «Элементы». Учитывая важность размеров атомов компонентов при образовании кристаллических структур различных типов, дополнительно было добавлено еще одно свойство: $(R_{covA} - R_{covB})/R_{covX}$, где R_{cov} — ковалентный радиус соответствующего элемента [208].

После анализа данных, в котором использовались несколько алгоритмов обучения ЭВМ [192], проводилось экзаменационное распознавание на материале обучающей выборки в двух режимах: без скользящего контроля и со скользящим контролем (табл. 7.11).

Для дальнейшего использования в коллективных методах принятия решения по результатам экзаменационного распознавания со скользящим контролем (табл. 7.11) были отобраны три алгоритма (выделены серым), показавших наилучшие результаты при решении задачи (линейный дискриминант Фишера, нейронные сети и k ближайших соседей).

Для прогноза возможности образования еще неполученных соединений состава AB_2X_4 использовались лучшие по результатам экзаменационного распознавания методы принятии коллективных решений: метод Байеса, метод логической коррекции и методы, основанные на нахождении шаблонов принятия решений и областей компетенции (табл. 7.12). Точность распознавания на уровне 99 %.

Таблица 7.10. Свойства элементов, использованные для описания соединений состава AB_2X_4

№	Свойство
1	Псевдопотенциальный радиус (по Цангеру)
2	Температура плавления
3	Ковалентный радиус
4	Квантовый номер
5	Расстояние до внутренних электронов (по Шуберту)
6	Расстояние до валентных электронов (по Шуберту)
7	Ионный радиус (по Бокию и Белову)
8	Температура кипения
9	Энтальпия испарения
10	Энтальпия плавления
11	Электроотрицательность (по Мартынову—Бацанову)
12	Энтальпия атомизации
13	Первый потенциал ионизации
14	Второй потенциал ионизации
15	Третий потенциал ионизации
16	Химический потенциал Мидемы (только для элементов A и B)
17	Номер группы (только для элементов A и B)
18	Регулярный номер (по Менделееву—Петтифору)
19	Температура Дебая (только для элементов A и B)
20	Молярная теплоемкость
21	Энтропия твердого тела
22	Теплопроводность
23	Количество валентных электронов (только для элементов A и B)

Таблица 7.11. Результаты экзаменационного распознавания объектов обучающей выборки с использованием различных алгоритмов распознавания образов (задача прогноза возможности образования соединений состава AB_2X_4)

Алгоритм	Оценка достоверности прогноза		Примечания
	со скользящим контролем, %	без скользящего контроля, %	
Алгоритм вычисления оценок	85.0	84.4	Все объекты 2 отнесены к классу 1
Двумерные линейные разделители	83.7	84.6	
Метод бинарных решающих деревьев	84.4	84.4	Все объекты 2 отнесены к классу 1
Линейный дискриминант Фишера	87.6	87.8	
Линейная машина	85.1	91.0	
Логические закономерности	83.5	91.4	
Логические закономерности			
Нейронные сети	86.7	90.3	
k ближайших соседей	88.0	99.9	
Метод опорных векторов	85.3	99.8	Почти все объекты 2 отнесены к классу 1
Голосование по тупиковым тестам	73.9	76.6	

Таблица 7.12. Результаты экзаменационного распознавания объектов обучающей выборки с использованием различных алгоритмов распознавания образов (задача прогноза возможности образования соединений состава AB_2X_4)

Алгоритм	Оценка достоверности прогноза, %
Метод Байеса	**99.9**
Области компетенции	**99.9**
Шаблоны принятия решений	**99.9**
Динамический метод Вуда	**95.8**
Выпуклый стабилизатор	**99.9**
Комплексный комитетный метод — голосование по большинству	**90.6**
Комплексный комитетный метод — усреднение	**90.6**
Логическая коррекция	**99.9**

Результаты прогноза по этим четырем методам сравнивались. Далее для прогнозируемых соединений прогнозировался тип кристаллической структуры при нормальных условиях. Табл. 7.14–7.22 содержат сводные результаты прогноза возможности образования и типа кристаллической структуры AB_2X_4 при мультиклассовом прогнозировании [191]. Т. е. одновременно рассматривались 17 классов (табл. 7.13):

Таблица 7.13. Условные обозначения типа кристаллической структуры AB_2X_4

Обозначение	Тип
1	структура шпинели
2	структура оливина
3	структура $MnEr_2S_4$
4	структура $CdAl_2S_4$
5	структура $PbGa_2Se_4$
6	структура β-K_2SO_4
7	структура $CaFe_2O_4$
8	структура Th_3P_4
9	структура $NiCr_2S_4$
10	структура $CaHo_2Se_4$
11	структура $PbBi_2S_4$
12	структура $GeSr_2S_4$
13	структура TlSe
14	структура $PbCr_2S_4$
15	структура сфалерита
16	другая структура (не 1–15)
17	отсутствия образования при н.у
<пусто>	неопределенность

Знаком «#»отмечены данные, использованные для обучения.

Таблица 7.14. Прогноз типа кристаллической структуры соединений состава $A^{IV}B^{II}{}_2S_4$ при нормальных условиях

A / B	Si	Ti	Cr	Mn	Ni	Ge	Zr	Nb	Mo	Rh	Sn	W	Re	Pb	U
Be			1			15	14	5	#1	1	15		1		15
Mg	#2	2	2			#2	2	16	6	1	#2	6	1		2
Ca	#2					#2					#2				
Ti	2		#9		#9	2	14	9		1	14		1	14	14
V	2	9	#9	#9	#9	2	14	9		1	14		1	14	14
Cr	14	#9		#1	#9	14		14		1	#14			#14	
Mn	#2	9				#2		14			2			2	
Fe	#2	#9				#2	#17			#1	#16	#17		1	
Co	2		1	1	#1	2			1	#1	2			11	
Ni			1	1		2			1	#1	2		1	11	16
Cu	2		1	1	1	2	14			1	#1		1	11	14
Zn	15	15				#15	15	15				15		11	15
Ga	16	16	#1	#16	#16			5	1	1	16			#5	
Ge	2	16	11	11	16	2	16	16	11	1	16			11	
Sr	12		#3			#12								8	3
Pd	15		1	1	1	15	11	11		1	11		1	11	14
Ag			1	1	1					1	14		1	11	14
Cd	15	15					15	15	15	1	15	15		11	15
Sn		8				#17		16		1				11	
Ba	#6	#6	6	6	6	#6	#16	6	6			6	6	8	6
La	12	8	8								8			#8	3
Ce	12	8	8	12			8	8	8		8	8	12	#8	12
Pr	12	8	8	#17	8		8	8	8		8	8	8	#8	6
Nd	12	8	8	#17	8		8	8	8		8	8	8	#8	
Pm	12	8	8				8	8			8			8	
Sm	12	8	#8	#17	8		8	8	8		8	8	8	#8	#17
Eu	#12	8	8		12	#12	8		12			12	3	#8	
Gd	12		#8	#8		12		12			8		12	#8	17
Tb	12		#3	#3		12					8			#8	
Dy	12	3	#3	#3		12		3	3		8	3			
Ho	2	3	#3	#3		2		3	3			3		#7	
Er	2	3	#3	#3				3	3			3		#7	
Tm	2	3	#3			2		3	3			3		#7	
Yb	12	#3	#3	3		#12		3	3			3	3	#7	
Lu	2		#3	#1		2				1			1	#7	7
Pt	15	11	1		1	15	11	11		1	14		1	11	14
Au	15		1	1	1	15				1	14		1		14
Hg		15		11	15		15	15	15	1		15	15	11	15
Tl			3			16					16			#17	
Pb		16				#16	16	16			8		6		
Ra	6	6	6	6	6	6		6	6		8	6	6		6
U	2		8			2			16		8	16	16	8	
Np	12					12		8	8		8			8	
Am	2	8	8			2					8			8	

Таблица 7.15. Прогноз типа кристаллической структуры соединений состава $A^{IV}B^{II}{}_2Se_4$ при нормальных условиях

A / B	Si	Ti	Cr	Mn	Ni	Ge	Zr	Nb	Mo	Rh	Sn	W	Re	Pb	U
Be	2					2	16	16	9						5
Mg	#2		2	2	2	2			6		#2	6			
Ca	#2	2	2	2	2	2			6		2				
Ti			9	#9	#9			9	9			9	9		
V		9	#9	#9	#9		9	9	9			9	9	14	
Cr	#17	#9		9	#9	#17		9		1	14			14	
Mn	#2	9	9		9	2					2			2	
Fe		#9	9	9	9	2		9	9	1					
Co		9				2				1		17			
Ni	17					2									
Cu	17					17		17							
Zn		15	15			#15		15	15	15	#17	15		17	17
Ga	17			#4		#17					#17			#5	
Ge	17					2					#17			17	17
Sr	12	6		12	12	12	12	6	6	6	12	6	6		
Pd							17			1	17			17	
Ag	17	17	17	17	17	#17					#17			#17	17
Cd	#17										#17			17	17
Sn						#17									
Ba	6	6	6	6	6	#12		6	6	6		6	6		
La	12		#8	12	12	12			16	16	2	16	16	#8	12
Ce	12		#8	12	12	12	12		12	12	12	12	12	#8	12
Pr	12		#8	12		12		16	12					#8	
Nd	12		#8			12							12	#8	
Pm	12			9		12			12					8	
Sm	12		#8	16		12		16						#8	
Eu	12	12	#8	12	12	12			12	12	12	12	12	8	
Gd	12					12									
Tb	12		10			12									
Dy	12					12	10								
Ho	12					12	10								
Er	12					12	10	10						#7	
Tm	12					12								#7	
Yb	12	12		12	12	#12			12			12		#7	
Lu	2					2								#7	
Pt	17										17				
Au	17	17	17	17	17	17					#17			17	
Hg						#4					#4				
Tl											#17			17	
Pb											#17				
Ra	6	6	6	6	6	12		6	6	6	6	6	6		
U	12	9	9	9	9	12			9	9	2		9	8	
Np	12	9	9			12			9					8	
Am	12	9				12					2			8	

Таблица 7.16. Прогноз типа кристаллической структуры соединений состава $A^{IV}B^{II}{}_2Te_4$ при нормальных условиях

A / B	Si	Ti	Cr	Mn	Ni	Ge	Zr	Nb	Mo	Rh	Sn	W	Re	Pb	U
Be	2					2									
Mg	2			2	2	2			17		2				
Ca	2			2	2	2			17		2				
Ti			#9	9					9	9		9	9	14	
V		#9	#9	9	9				9	9		9	9	14	
Cr	17	#9		9	9	17								14	
Mn		9	9											#17	
Fe		9	9	9	9				9		17				
Co	17					#17					#17			#17	
Ni	17					17					17			#17	17
Cu	17	17	17	17	17	#17	17	17	17	17	#17	17	17	17	17
Zn		17	17		17		17	17	17	17	17	17	17	17	17
Ga	17			4	#17	#17					#17			#17	
Ge					#17					17				17	17
Sr	12	12			12	12	12		12		12	6	12		
Pd	17					17					17		17	17	17
Ag	17	17	17	17	17	#17	17	17	17	17	#17	17	17	#17	17
Cd	#17	17	17	17	17	#17	17	17	17	17	17	17	17	17	17
Sn	17		17	17		#17								17	
Ba	#12	12	6		16	12			6	6		6	6		
La	12			12	16	12	16		16	16			16		16
Ce	12		16	12	12	12	16	16	16	16		16	16	17	12
Pr	12		16	16	16	12		16	16	16			16		
Nd	12			16	16	12		16					16		
Pm	12		16	16	16	12	16	16	16	16			16		
Sm	12	16	16	16	16	12		16	16				16	17	
Eu	12	12		12	12	12			12		17			#17	
Gd	12					12				16			16		
Tb	12				16	12				16			16		
Dy	12			16	16	12			16	16			16		
Ho	12		3		16	12				16			16		
Er	12		3		16	12							16		
Tm	12			3	16	12			3	16		3	16		
Yb	12	12		12	12	12			12						
Lu	12		16	16	16	12			16	1		1	16		
Pt	17					17					17			17	17
Au	#17	17	17	17	17	17	17	17	17	17	#17	17	17	#17	17
Hg		17	17	17	17		17	17	17	17		17	17		17
Tl	17		17			#17					#17			#17	
Pb	#17			#17	17	#17				17	#17				17
Ra	12	6	6		12	12		6	6	6		6			
U	12		9	9		12			9	9	2		16		
Np	12					12							16		
Am	12			16	16	12								17	

Таблица 7.17. Прогноз типа кристаллической структуры соединений состава $A^{II}B^{III}_2S_4$ при нормальных условиях (часть 1)

A / B	Be	Mg	Ca	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	Sr	Pd	Ag	Cd	Sn	Ba
B	15	16	5			15	16	16	16	16	17	17		15		16	17	#17	17	#16
Al	2	#2	5	#1		#1	#16	16			#1	#1	2	2	#5			#4	2	#16
P	15	15	15		15	15	11	16		16	17	17	17	15	15		17		15	15
Sc	1	#1	#7			1	#1	#1	1	1	1	#1	1	2	#7		1	#1	7	7
Ti			14		9	#9		#9	#9	#9	#1	1		2	14		1	1	14	
V			14	9		#9	#9	#9	#9	#9	#1	1		14	#14		1	1	14	
Cr	1			#9	#9		#1	#1	#1	#9	#1	#1	#1	14	14	1	1	#1	#14	#14
Mn		1	2	9				1			1	#1		#2	14	1	1	1	2	14
Fe	1	1	5	#9	#9						1	1	1	#2		1	1	1	#16	#16
Co	1	1	14			1	1	1		#1	#1	1	1	2	14	1		1	11	14
Ni	1	1				1	1	#1	#1		1	1	1	2	14	1	1	1	2	11
Ga	16	#16	#5	16	5	#1	#16	#16	16	#16		#4		17	#5	16	15	#4	16	#16
As	15	15	11	11	11	11	11	16	16	16	17	17	#17	17	#15	15	#17	#17		11
Y		#3				#3	#3					#17	2	2	#7			#1		#7
Mo	14		14			#16			#9		1	1	1	2	14	1	1	1	14	14
Rh	1	1		9		1	1	#1	#1	#1	#1	1	1	14	14	1	1	1	14	14
In		#1	#1	16		#1	#1	#1	#1	#1		#16	17	#17	#5	1	1	#1		#5
Sb	15	11	11	16	16	15	11	#16	16	16		17	17	2	11	16	17	#17	16	
La	17	#8	#8	8	8					12		#17	8		#8	8	8	8	8	
Ce	17	17	#8	8	8				12	12	12	#17	8		#8	8	8	8	8	
Pr	17	17	#8	8	8		#17	12		8	8	#17	8		#8	8	8	8	8	
Nd	17	17	#8	8	8		#17	#17	17	8	8	17	8		#8	8	8	8	8	#7
Pm	17	17	8	8	8	8	17	17	17	17	8	17	8				8	8	8	7
Sm	12	17	#8	8	8	#8	#17	12		8	8	17	8		#8	8	8	8	8	#7
Eu	12		8	8	8	8	17		12	12	12	17	8	#12	8	8	8	8		7
Gd	17	#8	#8			#8	#8	#17				#17	8	12	#8		8	8	8	#7
Tb		#3	#8			#3	#3					#17	8	12	#7				8	#7
Dy		#3	#8	3	3	#3	#3	#3						12	#7					#7
Ho	3	#3	#8	3	3	#3	#3	#3				#17		2	#7			#1		#7
Er		#3		3	3	#3	#3	#3							#7			#1		#7
Tm				3	3	#3		#3				#2	2	2	#7			#1		#7
Yb				3	#3	#3	3	#3	3			#2	12	#12	#7			#1		#7
Lu		#1				#3	#1	#1			1	#2	2	2	#7	1	1	#1		#7
Ir	1	1	15	9		1	1	#1	#1	#1	#1	1	1	2	14	1	1	1	14	14
Au	1	1	15			1	1	1	1	1	1	1		15		1	1	1	14	
Tl		16										17	17	16					16	
Bi		16	11		17	17	17	#17	17			#17	#17	#17	#16		#17		11	
Th	12	14	8	8	8	8	16	16				17	8	12	14	16	8		8	14
Pa	12	14	8			8					16		8	12	8	8	8	8	8	14
U	2	16	8			8	16	16		16	8	17	8	2	8	#16	8		8	
Np			8			8						17	8	12	8		8		8	
Pu	17	17	8			8	17	17			15		8	12	8	8	8	8	8	
Am	17	17	8	8	8	8	17	17					2	2	8	8	8		8	

Таблица 7.18. Прогноз типа кристаллической структуры соединений состава $A^{II}B^{III}{}_2S_4$ при нормальных условиях (часть 2)

A / B	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Pt	Au	Hg	Tl	Pb	Ra	Np	Am
B	5	5	5	15	5	5	5	5	5	5	5	5	5	5	5	15	17	17		17		5	5
Al	5	5	5	5	5	#5	#5	5	5	5	5	5	5	#5	5			#4	5	#5	5	5	5
P	15	14	15	14	15	11	11	14	15	15	15	5	11	5	15	15	17	4	17	11	15	14	11
Sc	7	7	7	7	7		#7	7		2	2	2					1			#7	7	7	7
Ti	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14			1	14	14		14	14
V	14	14	14	14	14	14	#14	14	14	14	14	14	14	14	14			1	14	14		14	14
Cr	14	14	14	14	14	14	#14	14	14	14	14	14	14	#14	14	1	1	#1		#14	14	14	14
Mn	2	14	14	2	2	11	14	2	2	2	2	2	2	14	17	1	1	1		2	14	14	2
Fe	17			2	11			9	9	9	9	9	9		17	1	1	1		1	16	1	11
Co	17	14	14	14	14	11	11	14	14	14	14	14	11		17	1	1	1		11	14	14	14
Ni	14	14	14	16	11	11	11	16	16	16	16	16	11		16	1	1	1		11	14	16	11
Ga				17		#5	#5	5					5	#5		16	15	#4	5	#5	5		
As	15	15	15	15	11	11	11	11	11	11	11	11	11	#16	15	15		#17	11	#11		11	11
Y	7	7	7			#7	#7												7	#7			3
Mo	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	1	1	1	14	14	14	14	14
Rh	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	1	1	1	14	14	14	14	14
In	17		5		5	#5	#5	5	5	5	5	5	5	#5	5	1	1	#1	16	16	5	16	5
Sb	15	15	15	15	11	11	#11	16	16	16	16	16	11	#11	16	15		#17	11	11	11	11	11
La		#8	8	8	8	#8	#8	#16	8	8	8	8	8	#8	8	3	8	8	8	#8		8	8
Ce	8		8	8	8	#8	#8	8	8	8	8	8	8	#8	8	8	8	8	8	#8	8	8	8
Pr	8	#8		8	8	#8	#8	8	8	8	8	8	8	#8	8	8	8	8	8	#8	8	8	8
Nd	8	#8	8		8	#8	#8	8	8	8	8	8	8	#8	8	8	8	8	8	#8		8	8
Pm	8	8	8	8		8	8	8	8	8	8	8	8	8	8		8	8	8			17	8
Sm	#8	#8	8	8	8		#8	8	8	8	8	8	8	#8	8	8	8	8	8	#8		17	8
Eu	#8	#8	8	8	8	#8		8	8	8	8	8	8	#8	8	8	8	8	8	#8		17	8
Gd	#16	8	8	8	8	#8	#8		8	8	8	8	8	#8	8		8	8	8	#8		8	8
Tb	8	8	8	8	8	#8				8	8	8	8	#8				8	8	#8	7	3	8
Dy			8	8	8	#8			8		8	8	8	#8							7	17	8
Ho	7	7	7		3	#7	#7								3					#7	7	17	
Er	7	7	7		3	#7	#7							#3	3					#7	7	17	
Tm	7	7	7	7		#7	#7	7						#3	3					#7	7		
Yb	7	7	7		3	#7	#7													#7	7	17	
Lu	7	7	7			#7	#7		2	2	2	2					1			#7	7	7	7
Ir	14	14	14	14	11	14	11	14	11	11	14	14	11	14	14	1	1	1	11	14	14	14	11
Au	15		14	14	8	11	11	14	15	15	15	15	11	14	14	1		1	11		15	7	8
Tl						5	5													#17	7		
Bi	11				11	11	#11		11	11	11	11	11	#16			17		11	#11	16		11
Th	14	8	8	8	8	8	8	14	8	8	8	8	8	8	14	16	8	8	8	8	14	14	8
Pa	8	8	8	8	8	8	8	14	8	8	8	8	8	8	8	16	8	8	8	8	14	14	8
U	8	8	8	8	8	8	8	16	8	8	8	8	8	#8	8	16	8	8	8	8	8	8	8
Np	8	8	8	8	8	8	8		8	8	8	8	8	8	8		8		8	8	8		8
Pu	8	8	8	8	8	8	8		8	8	8	8	8	8	8	8	8	8	8	8	8	17	8
Am	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	17	

Таблица 7.19. Прогноз типа кристаллической структуры соединений состава $A^{II}B^{III}_2Se_4$ при нормальных условиях (часть 1)

A / B	Be	Mg	Ca	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	Sr	Pd	Ag	Cd	Sn	Ba
B	16		5								17	17		17	5	17	17	#17	17	
Al		#16	#5				4		4			#4	17		#5			#4		#16
P	16	16	5												5		17			14
Sc	2	#1					#1					1		2	7			1		7
Ti	9				#9	9	#9	#9	#9	#9	1	1				9	1	1		7
V	9			9		#9	#9	#9	#9	#9	1	1			#14	9	1	1		
Cr		1		#9	9		9	9	9	#9	#1	#1		#17	14		1	#1	14	14
Mn	2	1		9	9	9		9	9	9	#1	#1			2	9	1	1	2	2
Fe		1		#9	#9	9	9		9	9	1	1		2		9	1	1		16
Co		1		9	9	9		9		9	1	1					1	1		16
Ni		1		9	9						1	1					1	1	17	16
Ga		16	#5				#4				#15	#4		#17	#5	17	15	#4	#17	#5
As			11								17	#17	#17	#17	16		#17	#17		11
Y	2	#1	#10		9	9		9	9	9		1	10	2	#7			#1		#7
Mo		1		9	9	9	9	9	9	9	1	1					1	1	14	
Rh	16	1		9	9	#9			#9	#9	#1	1		15			1	1		
In	17	#16	#16				#16					#16	17	#17	16	17	17	4	#17	16
Sb	16	16	16								17		17		16		17			
La	12	8	8			#8	12		12	12			8	12	8	16		#8	8	
Ce	12	8	8			#8	12	12	12	12			12	12	8	12		8		
Pr	12	8	8			#8					17			12	8	16	17	8	8	
Nd	12	8	8			#8	16				#17			12	8	16	17	#8	#16	7
Pm	2	8	8			8					17			2		16	17	8	2	7
Sm	12	8	8	16		#8					17			12	8	16	17	8		#7
Eu	12	8		12		#8	12	12	12	12				12	8	12		8	12	7
Gd		8									#17			12	8			#8		#7
Tb		16					10							2	#7			#8		7
Dy		1	#10									1		2	#7			#1		#7
Ho		#1	#10									1		2	#7			#1		7
Er		#1	#10									1		2	#7			#1		7
Tm	2	#1	10						2			1		2	#7			#1		7
Yb	12	#1	#10	12				12	12	12		1		#12	#7	12		#1		#7
Lu	2	#1	#10	10					2			1		2	#7			#1		#7
Ir		1	11								1	1					1	1		
Au		1	17	17	17	17	17			17	17		17	17		17			#17	
Tl		16	10								17		17		16	17	#17		#17	
Bi		16	16								17	#17	17	16	#16	17	17			
Th	12	8		9	9	9	9	9	9	9				12	8		17			
Pa	12	8		9	9	9	9	9	9	9				12	8		17			
U	16	16			9		9	9	9					2	8	16	17		2	
Np	12				9		9	9	9			1		12	8	16			2	
Pu	12	8		9	9		9	9	9	9		16		12	8	16		8		
Am	12	8		9	9						17			2	8		17		2	

Таблица 7.20. Прогноз типа кристаллической структуры соединений состава $A^{II}B^{III}{}_2Se_4$ при нормальных условиях (часть 2)

A / B	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Pt	Au	Hg	Tl	Pb	Ra	U	Np	Am
B					5	5		5	5			5			5	17	17	17	17					5
Al	5	5	5		5	#5	#5	5	5	5	5	5	5	#5	5			#4	5	#5	5	5		5
P							5					5		5	5	17	17			14	14			
Sc		7												2						#7	7			
Ti																9	1	1			7			
V							#14									9	1	1	14	14			14	
Cr		14					#14							#14			1	#1	14	14	14		14	
Mn						2	2							2			1	1	1	2	2			
Fe																	1		1	17	16			
Co																	1	1	1	11	16			
Ni													16	5				1	1	17	16			
Ga	5	5	5		5	#5	#5	5	5	5	5	5	5	#5	5			#4		#5	5	5	5	5
As	17	17	17	#17	17	#17			17				16				17	#17	17	11		17	17	17
Y		7																		7	7			
Mo		14					11										1	1			14		14	
Rh							14					16					1	1	1	14	14		14	
In						#5	#16						16	#5		17	17	#4		#16	16			
Sb	17	#17	17	17		11	#11		17		16	16	#16		16		17	#17	17	11	11	17	17	16
La		8	8	8	8	8	#8	16	8	8	8	8	8	16	16			8	8	#8			8	8
Ce	12		12	8	8	8	#8	12	8	8	8	8			12	12		8	8	#8		12	8	12
Pr	8	8		8	8	8	#8	16	8	8	8	8		#17	8			8	8	#8			8	8
Nd	8	8	8		8	#8	#8	16	8	8	8	8		#17	16		17	8	8	#8	7		8	8
Pm				8		8	8		8	8	8	8			8		17	8	8		7		17	8
Sm	8			8	8		#8		8	8	8	8		#17	8		17	8	8	#8	7		8	8
Eu						8										12		8	8	8	7			
Gd							8							#17		10		8	8	8	7			
Tb							#8							#17		10		8			7			
Dy									10		10	10		#17		10				7	7			
Ho		7		10	10		#7	10	10	10		10	10		10					7	7			
Er		7		10	10		#7	10	10	10	10		10		10					#7	7			
Tm		7					#7		10	10	10	10			10					#7	7			
Yb		7		10			#7	10	10	10	10	10	10		10					#7	7			
Lu		7		10	10		#7	10	10	10	10	10	10							#7	7			
Ir							11	16	11	11	11	11	16				1	1						
Au	17	17	17	17	17			17	17	17	17	17			17	17			17			17	17	17
Tl	17	17	17	17		5	5									17	17	4		17		17	17	10
Bi	17	#17	17	17		11	11		17		16		16	5	16	17	#17	#17	17	#10		17	17	16
Th						8	8										17	8	8	8				
Pa						8	8										17	8	8	8				
U					8	8	8											8	8	8			8	
Np						8	8											8	8	8				
Pu					8	8	8											8	8	8				
Am						8	8										17	8	8	8				

Таблица 7.21. Прогноз типа кристаллической структуры соединений состава $A^{II}B^{III}_2Te_4$ при нормальных условиях (часть 1)

A / B	Be	Mg	Ca	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	Sr	Pd	Ag	Cd	Sn	Ba
B		16										17				17	17	17		
Al		#16					4			17	17	#4	17		#13	17	17	4	17	#13
P		16									17			16	13	17	17		17	13
Sc	2	1										1		2	7		17	1		7
Ti	9				#9	#9	9	9	9		1	1			13			1		
V	9			#9		#9	9	#9	9	9	1	1			13	9				
Cr				#9	#9		9	9	#9	9	#1	#1		17	13		1	#1	14	13
Mn				9	9	9		9	9			#1			13			1	17	13
Fe				9	9	9	9		9	9		1		17	13				17	13
Co			13											#17	13				#17	13
Ni		16											17	17			17	#17	17	
Ga		#16	13				4			#17	17	#4		#17	13	17	#17	#4	#17	#13
As			13								17	#17			13	17	17	#17	#17	13
Y	2		#10											2	7					#7
Mo			10	9	9	9	9	9	9	9	1	1		16	13	16		1		13
Rh			13			#9			#16	#16	1	1			13	16	1		17	13
In		#16	13			#17	#4			17	17	#4	17	#17	13	17	17	#4	17	13
Sb			13		17	17			#17	17	17	17		#16	13	17	17			13
La	12		#17			17	12							12		16		#8	16	
Ce	12		#17				12		12	12	12			12		12	12	#8	12	13
Pr	12		#17	16			16							12		16		#8	16	
Nd			#17				16					8		12		16		#8	16	
Pm							16							2		16		8		7
Sm	12		#17	16			16							12	16	16		#8	16	#7
Eu	12			12			12	12	12	12	12			12	13	12	12	#8		
Gd		16												12	16	16		#8	16	#7
Tb		16					16							12		16		8		#7
Dy			#10				16							12	7	16				#7
Ho			#10				16							2	7	16				#7
Er			#10				16							2	7					#7
Tm	2		#10				16							2	7					#7
Yb	12		#10	12			12	12	12	12				12	7	12				7
Lu	2		#10				16			16				2	7					7
Ir	16		13										17	16	13	16			17	13
Au	17			17	17	17	17	17	17	17	17		17	17		17	#17		#17	
Tl						17					17		17	#17	13	17	17		#17	13
Bi		16	13		17	17	#17	#17	17		#17	#17	#17	#16	13			#17		13
Th	12			16	9	9	9	9	9			8		12		16			16	
Pa	12	16			9	9	9	9	9	9				12	14	16			16	
U	16				9	9	9	9	9		15	4		16	16	16		8	16	
Np	12				9	9	9	9	9			17		12	13	16	12		16	13
Pu	12						9							12	13			8		13
Am	12						16							2	13	16				13

Таблица 7.22. Прогноз типа кристаллической структуры соединений состава $A^{II}B^{III}{}_2Te_4$ при нормальных условиях (часть 2)

A / B	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	Pt	Au	Hg	Tl	Pb	Ra	U	Np	Am
B						5	5							5		17	17	17						
Al						#5	#5						5	#5		17	17	#4	17		13			
P	13					5								5	13	17	17		17	14	13			
Sc							16							16			17			7	7			
Ti							14										17			14				
V							14							15		9			14	14			14	
Cr		14					#14							#14				1	14	14	13		14	
Mn																		1	17	#17	13			
Fe																					13			
Co																			17	#17	13			
Ni	17			17															17	#17		17	17	17
Ga	13					#5	#5	5					5	#5	5	17	#17	#4	#17	#17	13			5
As		17						17							17	17	#17	#17	#17	#17	13	17	17	
Y							16							16					7		7			
Mo		13					14							16					14	14	13		14	
Rh		13	13	13	13	16	14	16	16	16	16	16	16	16	16	16					13			
In						#5	#5							#5		17	17	#4	17	#17	13			
Sb	17	#17	17	#17	17	#8	#8	17	#17			17	16	#8	17	#17	#17	#17	17	#17	13	17	17	17
La		16	16			8	16	16	16	16		16	16	8	16			8					8	8
Ce	12		12			12	8	12	12	12			12	8	12		12	8		8	13	12	8	12
Pr	8	16				8	16	16	16	16		16	16	16	16			8	8	16	13		8	8
Nd						8	16	16		16		16	16	16	16			8		16			8	8
Pm						8	16	16					16	16	16						7			
Sm							#16	16					16	16	16						7			
Eu														8			12	12		#17				
Gd							16							16						16	7			
Tb							16							16							7			
Dy							16				10	10	10	16	10					7	7			
Ho							#16			10		10	10	16	10					7	7			
Er							#16			10	10		10	16	10					7	7			
Tm							16			10	10	10		16	10						7			
Yb							16			10	10	10	10		10		12				7			
Lu							#16			10	10	10	10	16						7	7			
Ir		13	13	13	13		16	16	16	16	16	16	16	16	16	17			17	17	13	17		
Au	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17			17	#17		17	17	17
Tl	17					5	5					17				17	17	4		#17	13	17		
Bi	#17	17	17	#17	17	#8	#8	17	17	#17	17	17		#8	13		#17	#17	17	#16	13	17	17	17
Th							16							16						14				
Pa							16							16								8		
U						8	16							16										
Np						8	8							8							13			
Pu						8	16							8							13	17		
Am							16							16							13			

Таким образом, интеграция ИС СНВМ позволяет провести системный анализ больших массивов хранящейся информации с целью поиска взаимосвязей в данных и использования их для компьютерного конструирования новых веществ с заданными свойствами. Именно системный подход как комплексное взаимосвязанное последовательное рассмотрение всех факторов, путей и методов решения при поиске сложных взаимосвязей в данных послужил основой для создания информационно-аналитической системы для компьютерного конструирования неорганических соединений. В качестве информационной базы в нее вошла разработанная в настоящем исследовании интегрированная ИС СНВМ, а в качестве инструментальных средств анализа данных были использованы эмпирические методы распознавания образов. Полученные результаты прогнозирования свидетельствуют о перспективности применения интегрированной ИС СНВМ в качестве информационной основы для компьютерного конструирования неорганических соединений.

Краткие выводы

В главе проведены исследования по использованию разработанной интегрированной ИС СНВМ в качестве источника данных для системы компьютерного конструирования неорганических соединений и получены следующие результаты:

- Разработана методика применения интегрированной ИС СНВМ в программном комплексе компьютерного конструирования неорганических соединений.
- Разработан алгоритм обработки отсутствующих значений в обучающих выборках с учетом специфики предметной области — неорганического материаловедения.
- Получен прогноз образования еще не полученных перспективных полупроводниковых соединений состава ABX_2 (X = S, Se, Te). Сравнение прогнозов с экспериментальными данными показало, что ошибка прогнозирования составила менее 5 %.
- Выполнено прогнозирование ширины запрещенной зоны халькопиритов состава ABX_2 (X = S, Se, Te, N, P, As или Sb). Ошибка экзаменационного прогнозирования оказалась около 20 %, а при применении метода «голосования» результатов прогнозов с использованием коллектива алгоритмов удалось добиться правильных прогнозов.
- Проведено прогнозирование образования более сложных по составу соединений $A_2B_2(XO_4)_3$ с кристаллической структурой лангбейнита, перспективных для поиска новых пьезоэлектрических, сегнетоэлек-

трических, нелинейнооптических, электрооптических и люминесцентных веществ. Ошибка прогнозирования составила менее 30 %.

- Осуществлен прогноз образования перспективных сегнетоэлектриков A_2BCl_4. Ошибка прогнозирования при коллективном решении на уровне 15 %.
- Осуществлен прогноз образования и типов кристаллической структуры перспективных термоэлектриков AB_2X_4 (X = S, Se, Te). Ошибка прогнозирования при коллективном решении на уровне 1 %.
- Доказана перспективность применения интегрированных ИС СНВМ для прогнозирования свойств неорганических веществ.
- На основе анализа полученных результатов прогнозирования сделан вывод о перспективности использования разработанной интегрированной ИС СНВМ как информационной основы для программного комплекса компьютерного конструирования неорганических соединений.

Заключение

В работе получены следующие результаты:

- Формализована иерархия понятий, используемая в неорганической химии и материаловедении.
- Дано определение релевантной информации в контексте интегрированной ИС СНВМ на уровне неорганических веществ и кристаллических модификаций.
- Разработана методология интеграции ИС СНВМ, объединяющая преимущества известных методов интеграции.
- На основе системного анализа современных методов интеграции российских и зарубежных ИС предложена архитектура ИС СНВМ, обеспечивающей информационную поддержку компьютерного конструирования неорганических соединений.
- Разработана методика применения интегрированной ИС СНВМ в программном комплексе компьютерного конструирования неорганических соединений.
- Разработан и реализован алгоритм для обработки неопределенных значений в признаковых описаниях на основе метода «ближайших соседей».
- Разработана методика консолидации данных по свойствам неорганических веществ, особенностями которой являются применение хранилищ данных и методов виртуальной интеграции.
- Разработан и внедрен в Институте металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН) программный комплекс, реализующий интегрированную ИС СНВМ, объединяющий российские и зарубежные информационные ресурсы по свойствам неорганических веществ и материалов.
- Создана единая точка доступа пользователей к информации, консолидированной в рамках интегрированной ИС СНВМ.
- Разработана и реализована ИС «Кристалл» по свойствам акустооптических, электрооптических и нелинейнооптических веществ (русско- и англоязычные версии).
- Разработана и реализована ИС «Bandgap» по ширине запрещенной зоны неорганических веществ, используемых в электронике.
- Разработана и реализована ИС «IRIC» по информационным ресурсам в области неорганического материаловедения.
- Использование интегрированной ИС СНВМ для компьютерного конструирования неорганических соединений, применимых в электронике, показало, что средняя ошибка прогнозирования — около 20 %.

Литература

1. *Киселева Н. Н.* Компьютерное конструирование неорганических соединений: использование баз данных и методов искусственного интеллекта // Ин-т металлургии и материаловедения им. А. А. Байкова. М.: Наука, 2005. 289 с.
2. *Drago V. J.*, *Kaufman J. G.* Technical features of the chemical and materials property data network services on STN international // J. Chem. Inf. and Comput. Sci. 1993. V. 33. № 1. P. 46–51.
3. 10 years STN International. Databases in Science & Technology. 1993. FIZ Karlsruhe. STN Service Center Europe. 45 p.
4. *Belsky A.*, *Hellenbrandt M.*, *Karen V. L.*, *Luksch P.* New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design // Acta Crystallogr. 2002. V. B58. № 3. P. 364–369.
5. *Wood G. H.*, *Rodgers J. R.*, *Gough S. R.*, *Villars P.* CRYSTMET — The NRCC metals crystallographic data file // J. Res. NIST. 1996. V. 101. № 3. P. 205–215.
6. http://www.nist.gov/srd/nist3.htm
7. http://www.nist.gov/srd/nist15.htm
8. *Carr M. J.*, *Chambers W. F.*, *Melgaard D. et al*. NIST (Sandia) ICDD electron diffraction database: A database for phase identification by electron diffraction // J. Res. NIST. 1989. V. 94. № 1. P. 15–20.
9. http://www.npl.co.uk/npl/cmmt/mtdata/sgsub.html
10. http://www.nist.gov/srd/nist31.htm
11. http://www.nist.gov/srd/nist12.htm
12. http://physics.nist.gov/cgi-bin/AtData/main_asd
13. NIST/NRIM High Temperature Superconductors Database: Version 2.0, Standard Reference Data Program. National Institute of Standards and Technology. Gaithersburg, MD.
14. http://www.uni-konstanz.de/ZE/Bib/stn/nistcera.htm
15. *van Hove M. A.*, *Hermann K.*, *Watson P. R.* The NIST Surface Structure Database — SSD version 4 // Acta Crystallogr. 2002. V. B58. № 3. P. 338–342.
16. *Drago V. J.*, *Kaufman J. G.* Technical features of the chemical and materials property data network services on STN international // J. Chem. Inf. and Comput. Sci. 1993. V. 33. № 1. P. 46–51.
17. *Helter S. R.* NIST/EPA/MSDC mass spectral database, PC version 3.0 // J. Chem. Inf. and Comput. Sci. 1991. V. 31. № 2. P. 352–354.
18. *Ларичев О. И.*, *Мошкович Е. М.* Качественные методы принятия решений. М.: Наука, Физматлит. 1996.
19. *Кафаров В. В.*, *Дорохов И. Н.* Системный анализ процессов химической технологии. М.: Наука, 1976. 500 с.

20. *Волкова В. Н.*, *Денисов А. А.* Основы теории систем и системного анализа. СПб: СПбГТУ, 2001. 512 с.
21. *Corey E. J.*, *Wipke W. T.* Computer assisted design of complex organic synthesis // Science. 1969. V. 166. № 10 Oct. P. 178–192.
22. *Цирельсон В. Г.*, *Бобров М. Ф.*, *Апостолова Е. С.*, *Михайлюк А. И.* Лекции по квантовой химии. М: Изд-во РХТУ, 1998. 350 с.
23. *Берсукер И. Б.* Строение и свойства координационных соединений. Введение в теорию. Л: Химия, 1971. 312 с.
24. *Левин А. А.* Введение в квантовую химию твердого тела. М.: Химия, 1974. 237 с.
25. *Даркен Л. С.*, *Гурри Р. В.* Физическая химия металлов. М.: Металлургиздат, 1960. 583 с.
26. *Криват Б.*, *Макленнен Д.*, *Танг Ч.* Microsoft SQL Server 2008: Data Mining — интеллектуальный анализ данных. СПб.: BHV, 2009. 720 с.
27. *Розенблатт Ф.* Принципы нейродинамики. М.: Мир, 1965. 480 с.
28. *Вапник В. Н.*, *Червоненкис А. Я.* Теория распознавания образов. М.: Наука, 1974. 415 с.
29. *Андерсон Т.* Статистический анализ временных рядов. М.: Мир, 1976. 760 с.
30. *Ким Дж.-О.*, *Мьюллер Ч. У.*, *Клекка У. Р.*, *Олдендерфер М. С.*, *Блэшфилд Р. К.* Факторный, дискриминантный и кластерный анализ. Пер. с англ. М.: Финансы и статистика, 1989. 215 с.
31. *Дубров А. М.* Обработка статистических данных методом главных компонент. М.: Статистика, 1978. 135 с.
32. *Jardine N. Sibson R.* Mathematical Taxonomy. L.: John Wiley and Sons, 1971. 286 p.
33. *Fisher R.* The use of multiple measurements in taxonomic problems // Ann. Eugenics, 1936. V. 7. P. 179–188.
34. *Burges C. J. C.* A Tutorial on Support Vector Machines for Pattern Recognition // Data Mining and Knowledge Discovery. 1998. № 2. P. 121–167.
35. *Igelnik B.*, *Pao U.-H.*, Le*Clair S. R.*, *Shen C. Y.* The ensemble approach to neural-network learning and generalization // IEEE Trans. Neural Networks. 1999. V. 10. № 1. P. 19–30.
36. *Рязанов В. В.* Оптимальные коллективные решения в задачах распознавания и классификации: дисс. … д-ра ф-м.н. М. 1994.
37. *Brodie M.* The Grand Challenge of Information Technology — Invited talk, CAiSE-2002.
38. http://www.forresterresearch.com
39. *Дударев В. А.* Подходы к интеграции гетерогенных баз данных по свойствам неорганических веществ // Перспективные материалы. Спец. вып. ноябрь 2007. М.: Интерконтакт наука: С. 246–251.
40. *Lopatenko A.* National networks of science and technology information. Semantic Web based architecture for access to research and technology data — Proc. of EVA2001, Moscow.

41. *Imhoff C.* Understanding the Three E's of Integration EAI, EII and ETL // Intelligent Solutions, Inc. April 2005. http://www.intelsols.com
42. *Halevy A. et al*. Enterprise Information Integration: Successes, Challenges and Controversies — SIGACM-SIGMOD 2005. Baltimore, Maryland, USA.
43. http://www.w3.org/XML/Query
44. *Inmon W. H.* Building the Data Warehouse. N. Y.: John Wiley, 1992.
45. *Bitton D.* Why EII will not replace the data warehouse — SIGACM-SIGMOD 2005. Baltimore, Maryland, USA.
46. http://www.tpc.org
47. *Draper D.* The Nimble experience — SIGACM-SIGMOD 2005. Baltimore, Maryland, USA.
48. *Goh C. H.*, *Bressan S.*, *Madnick S. E.*, *Siegel M. D.* Context interchange: New features and formalisms for the intelligent integration of information // ACM Trans. on Information Systems. 1999. V. 17. P. 270–293.
49. *Bergamaschi S.*, *Castano S.*, *Vincini M.*, *Beneventano D.* Semantic integration of heterogeneous information sources // Data and Knowledge Engineering. 2001. V. 36. P. 215–249.
50. *Киселева Н. Н.*, *Дударев В. А.*, *Земсков В. С.* Интегрированная система баз данных по свойствам неорганических веществ и материалов // Теплофизические свойства веществ и материалов. Труды XII Российской конференции по теплофизическим свойствам веществ. М.: Интерконтакт наука, 2009. С. 139–142.
51. *Cali A.*, *Calvanese D.*, *Giacomo G. D.*, *Lenzerini M.*, *Naggar P.*, *Vernacotola F.* Ibis: Semantic data integration at work // Lecture Notes in Computer Science. V. 2681. 2003. P. 79–94.
52. *Levy A.*: Logic-based techniques in data integration / ed. J. Minker : Logic Based Artificial Intelligence. Norwell, MA: Kluwer Academic Publishers. 2000.
53. *Halevy A. Y.*: Answering queries using views. Very Large Databases. V.10. 2001. P. 270–294.
54. *Lenzerini M.*: Data integration: A theoretical perspective // 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems. 2002. P. 233–246.
55. *Manolescu I.*, *Florescu D.*, *Kossmann D.*: Answering xml queries on heterogeneous data sources. In: VLDB '01: Proceedings of the 27th International Conference on Very Large Data Bases. San Francisco, CA. USA. Morgan Kaufmann Publishers Inc. 2001. P. 241–250.
56. *Convent B.*: Unsolvable problems related to the view integration approach. Proc. of International Conference on Database Theory. 1986. P. 141–156.
57. *Lopatenko A.* Query answering under Exact View Assumption in Local As View Data Integration System. Proc. of EVA2001. Moscow.
58. *Bravo L.*, and *Bertossi L.* Disjunctive deductive databases for computing certain and consistent answers to queries from mediated data integration systems // J. Appl. Logic. V.3 2005. P.329-367.

59. *Imielienski T.*, Witold *Lipski J.* Incomplete information in relational databases // JACM. 1984. V. 31(4). P. 761–791.
60. *Kolaitis P.* Course on constraint satisfaction, complexity, and logic / ESSLLI. Vienna, Austria. 2003.
61. *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 2004: Semantic Interoperability and Integration.
62. *Масютин В. В.*, *Дударев В. А.* Системный анализ технологий интеграции гетерогенных баз данных // Материалы VII международной научно-практической конференции «Новейшие достижения европейской науки — 2011». Т. 34. Математика. София, 2011. С. 35–36.
63. *Бездушный А. Н.*, *Кулагин М. В.*, *Серебряков В. А.*, *Бездушный А. А.*, *Нестеренко А. К.*, *Сысоев Т. М.* Предложения по наборам метаданных для научных информационных ресурсов // Вычислительные технологии. 2005. Т. 10. № S1. С. 29–48.
64. http://www.crystalimpact.com/pauling
65. *Земсков В. С.*, *Киселева Н. Н.*, *Петухов В. В. и др*. Банк данных по фазовым диаграммам полупроводниковых систем // Изв. вузов. Материалы электронной техники. 1998. № 3. С. 17–23.
66. *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. 2001. № 4. С. 50–55.
67. *Земсков В. С.*, *Киселева Н. Н.*, *Киселев Н. Н. и др*. Банк данных по фазовым диаграммам полупроводниковых систем «ДИАГРАММА» // Неорган. материалы. 1995. Т. 31. № 9. С. 1198–1203.
68. *Кравченко Н. В.*, *Бурханов Г. С.*, *Киселева Н. Н. и др*. Банк данных по свойствам кристаллов для управления лазерным излучением // Изв. АН СССР. Неорган.материалы. 1991. Т. 27. № 1. С. 164–165.
69. *Юдина Н. В.*, *Петухов В. В.*, *Черемушкин Е. А. и др.* Банк данных по свойствам акустооптических, электрооптических и нелинейнооптических веществ // Кристаллография. 1996. Т. 41. № 2. С. 490–495.
70. *Дегтярев Ю. И.*, *Подбельский В. В.*, *Киселева Н. Н. и др*. База данных по свойствам кристаллов акустооптических, электрооптических и нелинейнооптических веществ, доступная из Internet // Изв. вузов. Материалы электронной техники. 1999. № 3. С. 35–40.
71. *Белокурова И. Н.*, *Дударев В. А.*, *Земсков В. С. и др*. Базы данных по материалам для электроники, доступные пользователям Интернета // Информационное общество. 2001. № 5. С. 24–27.
72. *Киселева Н. Н.*, *Прокошев И. В.*, *Дударев В. А.* и др. Система баз данных по материалам для электроники в сети Интернет // Неорган. материалы. 2004. Т. 42. № 3. С. 380–384.
73. https://webrech.fiz-karlsruhe.de/CATALOG/newcryst.html

74. *Tomaszewski P. E.* Structural phase transitions in crystals. I. Database // Phase Transit. 1992. V. 38. № 3. P. 127–220.
75. http://www.codata.org/databases/Materials.html
76. *Faber J.*, *Fawcett T.* The Powder Diffraction File: present and future // Acta Crystallogr. 2002. V. B58. № 3. P. 325–332.
77. http://www.cryst.ehu.es/icsdb/about.html
78. *Титов В. А.*, *Косяков В. И.*, *Кузнецов Ф. А.* Об организации информационного обеспечения работ по термодинамическому моделированию процессов технологии твердотельных устройств // Проблемы электронного материаловедения. Новосибирск: Наука, 1986. С. 8–16.
79. *Земсков В. С.*, *Кузнецов Ф. А.*, *Уфимцев В. Б.* Банки данных по полупроводниковым и другим материалам электронной техники и процессам их получения // Изв. вузов. Материалы электронной техники. 1998. № 3. С. 13–16.
80. *Магарилл С. А.*, *Борисов С. В.*, *Подберезская Н. В. и др*. Кристаллические структуры неорганических веществ — база количественных данных. Принципы построения и опыт эксплуатации // Ж. структур. химии. 1995. Т. 36. № 3. С. 559–563.
81. *Вертопрахов В. Н.*, *Доленко Т. Н.*, *Кучумов Б. М.* Фактография в электронном материаловедении. Новосибирск: Наука, 1988. 101 с.
82. *Allen F. H.*, *Davies J. E.*, *Galloy J. J.*, *et al*. The development of versions 3 and 4 of the Cambridge structural database system // J. Chem. Inf. and Comput. Sci. 1991. V. 31. № 2. P. 187–204.
83. *Гурвич Л. В.* ИВТАНТЕРМО — автоматизированная система данных о термодинамических свойствах веществ // Вестн. АН СССР. 1983. № 3. С. 54–65.
84. *Девятых Г. Г.*, *Ковалев И. Д.*, *Крылов В. А. и др*. Информационно-расчетная система «Высокочистые вещества и материалы» // Изв. вузов. Материалы электронной техники. 1998. № 3. С. 44–51.
85. *Cheynet B.* THERMODATA: an integrated thermodynamic and inorganic physico-chemical information system // CODATA Bull. 1985. № 58. P. 18–22.
86. https://webrech.fiz-karlsruhe.de/webrech/DBSS/trcthermoss.html
87. http://ultra.ippe.obninsk.ru:8097/nea_databank/dbsurvey.htm
88. http://www.cas.org/ONLINE/DBSS/jicsteplusss.html
89. *Westhaus U.*, *Droge T.*, *Sass R.* DETHERM — a thermophysical property database // Fluid Phase Equil. 1999. № 158–160. P. 429–435.
90. *Buck E.*, *Frankl E. M.* Gaps in the pure component experimental physical property data base // Chem. Eng. Progr. 1984. V. 80. № 3. P. 82–87.
91. *Palmer D. A.* DIPPR — an improbable success // AIChE Symp. Ser. 1990. V. 86. № 275. P. 1–4.
92. http://www.nist.gov/srd/nist17.htm
93. http://properties.nist.gov/fluidsci/semiprop
94. *Ho C. Y.*, *Li H. H.* Numerical databases on materials property data at CINDAS / Purdue University // J. Chem. Inf. and Comput. Sci. 1993. V. 33. № 1. P. 36–45.

95. https://cindasdata.com
96. *Трусов Б. Г.*, *Стрельцов Ф. Н.*, *Огнивов В. В.* Использование автоматизированной системы термодинамических расчетов в технологических исследованиях // III Всесоюзн. конф. по проблемам получения и использования в народном хозяйстве данных о свойствах материалов и веществ. 25–27 августа 1987. М.: изд-во Госстандарта. 1987. С. 185–187.
97. *Spencer P. J.* Development of thermodynamic databases and their relevance for the solution of technical problems // Z. Metallk. 1996. Bd. 87. H. 7. S. 535–539.
98. *Dinsdale A. T.* SGTE data for pure elements // CALPHAD. 1991. V. 15. № 4. P. 317–425.
99. *Andersson J.-O.*, *Jansson B.*, *Sundman B.* THERMO-CALC: a data bank for equilibria and phase diagram calculations // CODATA Bull. 1985. № 58. P. 31–35.
100. *Kloffer M.* Gmelin-Online Datensystem. Ablauf der dezentralen Datenerfassung fur das Gmelin-Online-System von der Diskette zur Datenbank // Software-Entwickl. Chem. 1: Proc.Workshops Comput. Chem., Hochfilzen (Tirol), 19–21 Nov., 1986, Berlin e.a., 1987. P. 99–102.
101. *Vogt J.*, *Mez-Starck B.*, *Vogt N.*, *Hutter W.* MOGADOC — a database for gasphase molecular spectroscopy and structure // J. Mol. Struct. 1999. V. 485–486. № 1. P. 249–254.
102. http://www.kbk-sdi.com/sdisp/nel/ppds_window/PPDS_Window_Iss6. pdf
103. *Asada Y.*, *Nakada E.*, *Yokokawa T.*, *et al.* Database for materials design of multilayered superconductors // J. Mater. Sci. Soc. Jap. 1988. V. 24. № 4. P. 199–203.
104. *Chen H.*, *Iwata S.* Data analysis by a data system on high-Tc superconducting materials // Materials System. 1993. V. 12. Nov. P. 63–70.
105. http://www.cas.org/ONLINE/DBSS/asmdatass.html
106. http://www.cas.org/ONLINE/DBSS/copperdatass.html
107. http://www.bus.iastate.edu/mennecke/server/courses/RICDBSpecifications.htm
108. *Xu L.*, *Li G.*, *Wang S.*, et al. CIAC comprehensive information system of rare earths // J. Chem. Inf. and Comput. Sci. 1991. V. 31. № 3. P. 375–380.
109. http://theorie.physik.uni-wuerzburg.de/webrech/DBSS/mdfss.html
110. *Nakanomyo T.*, *Akiyama Y.*, *Itoh T.*, *et al*. Factual database on amorphous materials // Sci. Repts. Res. Inst. Tohoku Univ. 1992. V. 37. № 2. P. 228–236.
111. *Морозов Е. Г.*, *Ратнер И. М.*, *Авербух В. М. и др*. Реализация банка данных по люминофорам на персональном компьютере // Неорган. материалы. 1993. Т. 29. № 10. С. 1332–1337.
112. *Fokin L.*, *Popov V.*, *Kalashnikov A. et al.* Joint Russian and Bulgarian Academies of Sciences Database of intermolecular potentials and diffusion coefficients for components of the CVD processes in microelectronics // Int. J. Thermophysics. 2001. V. 22. № 5. P. 1497–1506.
113. *Акчурин Р. Х.*, *Берлинер Л. Б.* Информационно-расчетная система для компьютерного моделирования процессов жидкофазной эпитаксии // Изв. вузов. Материалы электронной техники. 1998. № 2. С. 51–56.

114. *Mishima Y.*, *Ishino S.*, *Iwata S.* An approach to information processing of phase diagrams // Mater. Sci. and Eng. 1973. V. 11. № 3. P. 163–176.
115. *Голикова М. С.*, *Бурханов Г. С.*, *Киселева Н. Н. и др*. Банк данных по свойствам акустооптических кристаллов неорганических соединений // Изв. АН СССР. Неорган. материалы. 1989. Т. 25. № 4. С. 700–701.
116. *Савицкий Е. М.*, *Киселева Н. Н.*, *Пищик Б. Н. и др*. Разработка автоматизированного банка данных по свойствам тройных неорганических фаз // Докл. АН СССР. 1984. Т. 279. № 3. С. 627–629.
117. *Киселева Н. Н.*, *Кравченко Н. В.* Банк данных по свойствам тройных неорганических соединений как основа для компьютерного конструирования новых веществ // Журн. неорган. химии. 1992. Т. 37. № 3. С. 698–702.
118. *Киселева Н. Н.*, *Кравченко Н. В.*, *Петухов В. В.* Банк данных по свойствам тройных неорганических соединений (вариант для IBM PC) // Неорган. материалы. 1996. Т. 32. № 5. С. 636–640.
119. Japan's first move towards collaborative ventures // III–Vs Rev. The Adv. Semicond. Mag. 2003. V. 16. № 8. P. 6.
120. *Ansara I.* Generation et application des bases de donnees thermochimiques // Entropie. 1991. V. 27. № 161. P. 74–79.
121. *Липаев В. В.* Технико-экономическое обоснование проектов сложных программных средств. М.: Синтег, 2004. 284 с.
122. The Fastest Webserver? 16 November 2011. http://www.webperformance.com/load-testing/blog/2011/11/what-is-the-fastest-webserver
123. Microsoft Windows Server 2003 with Internet Information Services (IIS) 6.0 vs. Linux Competitive Web Server Performance Comparison // Veritest report. April 2003. http://www.veritest.com
124. *Florian C.* Top most vulnerable applications and operating systems in 2010 // February 2011. http://www.gfi.com
125. Исследование на тему: Какая ОС безопаснее? Январь 2008. http://www.securitylab.ru
126. *Смагин В. А.*, *Солдатенко В. С.*, *Кузнецов В. В.* Моделирование и обеспечение надежности программных средств АСУ. СПб. 1999. 49 с.
127. Microsoft Windows Server 2003 vs. Red Hat Enterprise Linux AS 3.0: IT Professionals Running a Production Environment // Veritest report. April 2005. http://www.veritest.com
128. EICTA Interoperability White Paper. 21 June 2004. http://www.eicta.org
129. http://www.w3.org/2002/ws
130. *Wilcox J.*, *Sargent P.*, *Bayriamova Z.*, *Matiesanu C.* Interoperability: How Technology Managers Rate Microsoft and Its Technologies for Development // Jupiter Research (MIC04-C02). 7 April 2004. http://www.jupiterresearch.com
131. Roger Sessions. Interoperability Through Service-Oriented Architectures (SOAs). ObjectWatch. http://www.objectwatch.com
132. Magic Quadrant for Business Intelligence and Analytics Platforms. 20 February 2014. Gartner Research. http://www.gartner.com/technology/reprints.do?id = 1-1QLGACN&ct = 140210&st = sb

133. Linux vs. Windows: another fine Microsoft TCO Analysis. August 2008. http://www.zdnet.com
134. Server operating system licensing & support cost comparison Platforms. May 2009. http://www.bearingpoint.com
135. *Поляков А. Е.*, *Дударев В. А.* Хранилище данных для интеграции информационных систем по свойствам неорганических веществ // Интеграл. 2011. № 6. С. 18–19.
136. *Дударев В. А.* Программа удаленного администрирования базы данных по физико-химическим свойствам веществ / XXVIII Гагаринские чтения. Тезисы докладов Международной молодежной научной конференции. М.: МАТИ, 2002. С. 18–19.
137. *Дударев В. А.* Универсальная программа удаленного администрирования баз данных. Научный сервис в сети Интернет / Труды Всероссийской научной конференции. М.: МГУ, 2002. С. 75–77.
138. *Миано Дж.* Форматы и алгоритмы сжатия изображений в действии. М.: Триумф, 2003. 336 с.
139. *Дударев В. А.* Программа удаленного администрирования базы данных по свойствам кристаллов акустооптических, электрооптических и нелинейнооптических веществ. Новые информационные технологии / Тезисы докладов. Т. 2. М.: МГИЭМ, 2002. С. 359–360.
140. *Dudarev V. A.* Databases on properties of inorganic substances and materials integration infrastructure / Proceedings of the 3rd Asian Materials Database Symposium (AMDS 2012), 2012. P. 71–76.
141. *Wilson A. H.* The Theory of Metals. Cambridge, 1953.
142. *Регель А. Р.*, *Глазов В. М.* Периодический закон и физические свойства электронных расплавов. М.: Наука, 1978. 308 с.
143. *Косяков В. И.*, *Сурков Н. В.* Способы обработки и хранения информации о фазовых диаграммах // Геология и геофизика. 1998. Т. 39. № 9. С. 1192–1209.
144. *Масленков С. Б.*, *Удовский А. Л.* Банки данных по диаграммам состояний металлических систем: современное состояние, проблемы, перспективы их развития и применения / Расчеты и экспериментальные методы построения диаграмм состояния. М.: Наука, 1985. С. 77–87.
145. *Nash P.* Computer representation of phase diagrams // Bull. Alloy Phase Diagr. 1984. V. 5. № 1. P. 5–9.
146. *Кауфман Л.*, *Бернстейн Х.* Расчет диаграмм состояния с помощью ЭВМ. М.: Мир, 1972. 326 с.
147. *Новик Ф. С.*, *Кожевников И. Ю.*, *Гультяй И. И.* Банки данных по диаграммам состояния в автоматизированных системах научных исследований / Расчеты и экспериментальные методы построения диаграмм состояния. М.: Наука, 1985. С. 87–93.
148. *Murrey I.*, *Orser I.* Interactive computer graphics for storing of phase diagram // Bull. Alloy Phase Diagr. 1980. V. 1. № 1. P. 19–31.
149. *Дегтярев С. А.*, *Воронин Г. Ф.* Применение сплайнов в термодинамике растворов / Математические проблемы фазовых равновесий. Новосибирск: Наука, 1983. С. 53–83.

150. *Косяков В. И.*, *Малахов Д. В.* Принципы свертки и хранения информации о фазовых диаграммах / Прямые и обратные задачи химической термодинамики. Новосибирск: Наука, 1987. С. 73–80.
151. *Луцык В. И.*, *Воробьева В. П.*, *Сумкина О. Г.* Моделирование фазовых диаграмм четверных систем. Новосибирск: Наука, 1992. 199 с.
152. *Armstrong J.*, *de Haan J.* Macromedia Flash 8. Фирменное руководство. М.: Триумф, 2006. 256 с.
153. *Паркер Т.*, *Сиян К.* TCP/IP для профессионалов. СПб.: Питер, 2004. 864 с.
154. *Кришнамурти Б.*, *Рексфорд Дж.* Web-протоколы. Теория и практика. HTTP/1.1, взаимодействие протоколов, кэширование, измерение трафика. М.: Бином, 2002. 592 с.
155. *Киселева Н. Н.* Компьютерное конструирование неорганических соединений, перспективных для применения в электронике, с использованием баз данных и методов искусственного интеллекта: Дисс. … д-ра хим. наук. Москва, МИТХТ. 2004. 333 с.
156. *Дударев В. А.* Принципы интеграции БД по свойствам неорганических веществ и материалов / Теплофизические свойства веществ и материалов. Труды XII Российской конференции по теплофизическим свойствам веществ. М.: Интерконтакт наука, 2009. С. 128–132.
157. *Дударев В. А.* Интеграция информационных систем по свойствам неорганических веществ для информационной поддержки принятия решений при прогнозировании свойств веществ // Саарбрюккен: LAP Lambert Academic Publishing, 2012. 176 с.
158. *Масютин В. В.*, *Дударев В. А.* На пути к единой информационной системе по свойствам неорганических веществ // Интеграл. № 6. 2010. С. 30–31.
159. *Kornyshko V.*, *Dudarev V.* Software Development for Distributed System of Russian Databases on Electronics Materials // Information Theories & Applications. V. 13. № 2. 2006. P. 121–126.
160. *Дударев В. А.* Разработка общих стандартов и типового программного обеспечения для интеграции российских и зарубежных баз данных по свойствам неорганических веществ и материалов // Перспективные материалы. Спец. вып. Ноябрь 2008. М.: Интерконтакт наука: С. 174–179.
161. *Полянский А.* Программирование на CGI. М.: Майор, 2003. 176 с.
162. *Валиков А.* Технология XSLT. Спб.: БХВ-Петербург, 2002. 544 с.
163. *Биберштейн Н.*, *Боуз С.*, *Фиаммант М.*, *Джонс К.*, *Ша Р.* Компас в мире сервис-ориентированной архитектуры (SOA). М.: КУДИЦ-Пресс, 2007. 256 с.
164. *Орлик С. В.* Обзор технологических стандартов Web-служб и тенденций их развития / Сборник трудов II Всероссийской практической конференции «Стандарты в проектах современных информационных систем». М., 2002.
165. http://www.w3.org/XML/Schema
166. *Чаусов В. И.* Практическое применение языка XML в задаче интеграции бизнес-приложений / Сборник трудов II Всероссийской практической конференции «Стандарты в проектах современных информационных систем». М., 2002.

167. *Kiselyova N.*, *Iwata S.*, *Dudarev V.*, *Prokoshev I.*, *Khorbenko V.*, *Zemskov V.* Principles of integration of Russian and Japanese databases on inorganic materials / Fifth International Conference «Information Research, Applications» i.Tech 2007. Sofia: FOI ITHEA, 2007. V. 2. P. 326–333.
168. *Kiselyova N.*, *Iwata S.*, *Dudarev V.*, *Prokoshev I.*, *Khorbenko V.*, *Zemskov V.* Integration Principles of Russian and Japanese Databases on Inorganic Materials // Information Technologies and Knowledge. 2008. V. 2. № 4. P. 366–372.
169. *Дударев В. А.* Реализация интегрированной информационной системы, объединяющей Web-интерфейсы информационных систем по свойствам неорганических веществ и материалов. Руководство разработчика. http://meta.imet-db.ru/LinkIntegrationManual.doc
170. *Дударев В. А.*, *Киселева Н. Н.* Интеграция интерфейсов российских и японских баз данных по свойствам неорганических веществ / XIII российская конференция по теплофизическим свойствам веществ (с международным участием): Тезисы докладов. Новосибирск: Изд-во Института теплофизики СО РАН, 2011. С. 105–106.
171. *Дударев В. А.*, *Шмакова Е. Г.* Web-интерфейс для доступа к гетерогенным информационным системам по свойствам неорганических веществ // Интеграл. 2013. № 4. С. 55.
172. *Дударев В. А.*, *Филоретова О. А.* Подход к интеграции баз данных по свойствам неорганических веществ на основе метабазы // Прикладная информатика. 2013. № 4(46). С. 38–42.
173. *Дударев В. А.* XML-schema стандартизирующая формат XML-документа для обновления метабазы. http://meta.imet-db.ru/MUService.xsd
174. *Дударев В. А.* WSDL-контракт, оговаривающий методы взаимодействия с Web-сервисом обновления метабазы. http://meta.imet-db.ru/MUService/MUService.asmx?wsdl
175. Data Encryption Standard, National Institute of Standards and Technology. Federal Information Processing Standard (FIPS) Publication 46–1. Supersedes FIPS Publication 46, (January, 1977; reaffirmed January, 1988).
176. *Дударев В. А.* WSDL-контракт, оговаривающий методы взаимодействия с Web-сервисом обслуживающим интегрируемые ресурсы. http://meta.imet-db.ru/Service/Service.asmx?wsdl
177. *Дударев В. А.* XML-schema стандартизирующая формат XML-документа, возвращаемого Web-сервисом поиска релевантной информации, со списком релевантной информации. http://meta.imet-db.ru/Relevance.xsd
178. *Дударев В. А.* WSDL-контракт, оговаривающий методы взаимодействия с программными адаптерами интегрируемых информационных систем выполненными в виде Web-сервисов. http://crystal.imet-db.ru/EII_Crystal/EII_Crystal.asmx?WSDL
179. *Дударев В. А.* WSDL-контракт, оговаривающий методы взаимодействия с предметным посредником интегрированной информационной системы. http://meta.imet-db.ru/EII/Service.asmx?WSDL

180. *Kiselyova N. N.*, *Dudarev V. A.* Integrated System of Databases on Properties of Inorganic Materials and Substances / 2nd Asian Materials Database Sysposium. 2010. P. 3–4.
181. *Киселева Н. Н.*, *Дударев В. А.*, *Земсков В. С.* Интегрированная система баз данных по свойствам материалов для электроники / Теплофизические свойства веществ и материалов. Тезисы докладов XII российской конференции по теплофизическим свойствам веществ. М.: Интерконтакт Наука, 2008. С. 185–186.
182. *Дударев В. А.* Применение интегрированной системы баз данных для поиска новых полупроводниковых соединений / IX Российская ежегодная конференция молодых научных сотрудников и аспирантов «Физикохимия и технология неорганических материалов». Сборник материалов. М.: ИМЕТ РАН, 2012. С. 133–134.
183. *Киселева Н. Н.*, *Дударев В. А.*, *Столяренко А. В.*, *Земсков В. С.* Компьютерное конструирование неорганических соединений, перспективных для поиска новых материалов для электроники // Изв. вузов. Материалы электронной техники. 2006. № 3 С. 61–68.
184. *Дуда Р.*, *Харт П.* Распознавание образов и анализ сцен / Пер. с англ. М.: Мир, 1976. 507 с.
185. *Брюхов Д. О.*, *Вовченко А. Е.*, *Захаров В. Н.*, *Желенкова О. П.*, *Калиниченко Л. А.*, *Мартынов Д. О.*, *Скворцов Н. А.*, *Ступников С. А.* Архитектура промежуточного слоя предметных посредников для решения задач над множеством интегрируемых неоднородных распределенных информационных ресурсов в гибридной грид-инфраструктуре виртуальных обсерваторий // Информатика и ее применения. 2008. Т. 2. № 1. С. 2–34.
186. *Гладун В. П.* Процессы формирования новых знаний. София: СД Педагог 6. 1995. 192 с.
187. *Гладун В. П.* Партнерство с компьютером. Киев: Port-Royal, 2000. 119 с.
188. *Гладун В. П.* Растущие пирамидальные сети // Новости искусственного интеллекта. 2004. № 1. С. 25.
189. *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. Новосибирск: изд-во Ин-та математики, 1999. 269 с.
190. *Поляков Е. А.*, *Масютин В. В.*, *Дударев В. А.* Компьютерное конструирование неорганических соединений на основе интегрированной информационной системы // Прикладная информатика. 2012. № 4(40) С. 38–43.
191. *Дударев В.А.*, *Филоретова О.А.*, *Брыкина Г.В.* Использование методов распознавания образов для компьютерного конструирования неорганических соединений // Прикладная информатика, № 2(50), 2014, с. 82–87.
192. *Журавлев Ю. И.*, *Рязанов В. В.*, *Сенько О. В.* «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. 176 с.
193. *Шмакова Е. Г.*, *Поляков А. Е.*, *Дударев В. А.* Методика компьютерного эксперимента с целью поиска перспективных неорганических веществ // Технологии XXI века в легкой промышленности. 2013. № 7, часть 1, раздел 4.

194. *Журавлев Ю. И.*, *Никифоров В. В.* Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. № 3. С. 1–11.
195. *Журавлёв Ю. И.* Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1979. Т. 33. С. 5–68.
196. *Мазуров В. Д.* Комитеты системы неравенств и задача распознавания // Кибернетика. 1971. № 3.
197. *Растригин Л. А.*, *Эренштейн Р. Х.* Коллективные правила распознавания. М.: Энергия, 1981. С. 244.
198. *Larin S. B.*, *Ryazanov V. V.* The Search of Precedent-Based Logical Regularities for Recognition and Data Analysis Problems // Pattern Recognition and Image Analysis. 1997. V. 7. № 3. P. 322–333.
199. *Ryazanov V. V.* Recognition Algorithms Based on Local Optimality Criteria// Pattern Recognition and Image Analysis. 1994. V. 4. № 2. P. 98–109.
200. *Freund Y.*, *Schapire R.* A decision-theoretic generalization of online learning and an application to boosting / European Conference on Computational Learning Theory. 1995. P. 23–37.
201. *Freund Y.* Boosting a weak learning algorithm by majority / COLT: Proceedings of the Workshop on Computational Learning Theory. Morgan Kaufmann Publishers, 1990.
202. *Freund Y.*, *Schapire R.* Experiments with a new boosting algorithm / International Conference on Machine Learning. 1996. P. 148–156.
203. *Breiman L.* Bagging predictors // Machine Learning. 1996. V. 24. № 2. P. 123–140.
204. *Breiman L.* Bias, variance, and arcing classifiers / Tech. Rep. 460: Statistics Department, University of California, 1996.
205. *Skurichina M.*, *Kuncheva L.*, *Duin R.* Bagging and boosting for the nearest mean classifier: Effects of sample size on diversity and accuracy / Third International Workshop Multiple Classifier Systems, Cagliari, Italy. Ed. J. K. F. Roli. Berlin: Springer. 2002. V. 2364. P. 62–71.
206. *Kuncheva L.* Combining Pattern Classifiers: Methods and Algorithms. Wiley, 2004. 350 p.
207. *Ветров Д. П.*, *Кропотов Д. А.* Алгоритмы выбора моделей и построения коллективных решений в задачах классификации, основанные на принципе устойчивости. М.: КомКнига/URSS, 2006. 112 с.
208. *Масютин В. В.*, *Дударев В. А.* Компьютерное конструирование новых неорганических соединений состава AB_2X_4 // Материалы научно-технической конференции «Наукоемкие химические технологии 2011». МИТХТ им. М. В. Ломоносова. Москва, 2011. С. 24–25.
209. *Иоффе А. Ф.* Физика полупроводников. М.-Л.: изд-во АН СССР, 1957. 491 с.
210. *Рез И. С.*, *Поплавко Ю. М.* Диэлектрики. Основные свойства и применения в электронике. М.: Радио и связь, 1989. 288 с.
211. *Ярив А.*, *Юх П.* Оптические волны в кристаллах. М.: Мир, 1987. 616 с.

212. *Солодовников С. Ф.*, *Клевцова Р. Ф.*, *Клевцов П. В.* Взаимосвязь строения и некоторых физических свойств двойных молибдатов (вольфраматов) одно- и двухвалентных металлов // Ж. структ. химии. 1994. Т. 35. № 6. С. 145–157.
213. *Антоненко А. М.*, *Поздеев В. Г.* Упругие и пьезоэлектрические свойства $K_2Cd_2(SO_4)_3$ // Физ. тверд. тела. 1981. Т. 23. № 8. С. 2494–2496.
214. *Дударев В. А.*, *Воякин А. С.* Data Entity Framework. Свидетельство о государственной регистрации программы ЭВМ № 2010615354 от 20.08.2010 г.
215. *Дударев В. А.* Единая точка входа в информационные системы по свойствам неорганических веществ / X российская ежегодная конференция молодых научных сотрудников и аспирантов «Физико-химия и технология неорганических материалов». Сборник материалов. М.: ИМЕТ РАН, 2013. С. 84–86.
216. http://www.nist.gov
217. http://www.codata.org
218. http://www.stn.com
219. http://mits.nims.go.jp/db_top_eng.htm
220. *Belov G. V.*, *Iorish V. S.*, *Yungman V. S.* IVTANTHERMO for Windows — database on thermodynamic properties and related software // CALPHAD. 1999. V. 23. P. 173–180.
221. *Трахтенгерц М. С.* CDS/ISIS for Windows — новый эффективный инструмент для текстовых баз данных. Опыт Теплофизического центра ИВТ РАН // НТИ. Сер. 2. Информ. процессы и системы. 2006. № 6. С. 30.
222. *Трахтенгерц М. С.* Технология подготовки информации для баз данных в обменном формате ISO 2709 // НТИ. Сер. 2. Информ. процессы и системы. 2006. № 7. Р. 28.
223. *Зицерман В. Ю.*, *Кобзев Г. А.*, *Фокин Л. Р.* Информационный конвейер для справочных данных о теплофизических свойствах веществ / Теплогидравлические аспекты безопасности АЭУ с реакторами на быстрых нейтронах. Межотраслевая тематическая конференция. Тезисы докладов. ФЭИ. Обнинск, 87. 2005.
224. *Wang P.*, *Neumann D. B.* A Database and Retrieval System for the NBS Tables of Chemical Thermodynamic Properties // J. Chem. Inf. and Comput. Sci. 1989. V.29.P. 31.
225. NIST Standard Reference Database 88. NIST/TRC Ideal Gas Database. Version 2.0 .Users' Guide. NIST, Gaithersburg, 2006.
226. *Kaufman J. G.*, *Drago V. J.* Direct access to material properties for modeling and simulation // Modelling Simul. Mater. Sci. Eng. 1993. № 1. P. 335.
227. *Davies R. H.*, *Dinsdale A. T.*, *Gisby J. A.*, *Robinson J. A. J.*, *Martin S. M.* MTDATA — Thermodynamic and Phase Equilibrium Software from the National Physical Laboratory // CALPHAD. 2002. V. 26. P. 229.
228. *Huang Z.*, *Conway P. P.*, *Thomson R. C.*, *Dinsdale A. T.*, *Robinson J. A. J.* A computational interface for thermodynamic calculations software MTDATA // CALPHAD. 2008. V. 32. P. 129.

229. *Ho C. Y.*, *Li H. H.* Numerical databases on materials property data at CINDAS / Purdue University // J. Chem. Inf. and Comput. Sci. 1993. V. 33. P. 36.
230. *Sundman B.*, *Jansson B.*, *Andersson J.-O.* The THERMO-CALC Databank system // CALPHAD. 1985. V. 9. P. 153.
231. *Andersson J.-O.*, *Helander T.*, *Hoglund L.*, *Shi P.*, *Sundman B.* THERMO-CALC & DICTRA, Computational Tools ForMaterials Science // CALPHAD. 2002. V. 26. P. 273.
232. *Yokokawa H.*, *Yamauchi S.*, *Matsumoto T.* Thermodynamic database MALT2 and its applications to high temperature materials chemistry // Thermochim. Acta. 1994. V. 245. P. 45.
233. *Yokokawa H.*, *Yamauchi S.*, *Matsumoto T.* Thermodynamic Database MALT for Windows with gem and CHD // CALPHAD. 2002. V. 26. P. 155.
234. *Bale C. W.*, *Chartrand P.*, *Degterov S. A.*, *Eriksson G.*, *Hack K.*, *Mahfoud R. Ben*, *Melanqon J.*, *Pelton A. D.*, *Petersen S.* FactSage Thermochemical Software and Databases // CALPHAD. 2002. V. 26. P. 189.
235. *Ohnuma I.*, *Liu X. J.*, *Ohtani H.*, *Ishida K.* Thermodynamic database for phase diagrams of micro-soldering alloys // J. Electron. Mater. 1999. V. 28. P. 1164.
236. *Liu X. J.*, *Ohnuma I.*, *Wang C. P.*, *Kainuma R.*, *Ishida K.*, *Ode M.*, *Koyama T.*, *Onodera H.*, *Suzuki T.* Thermodynamic database on microsolders and copper-based alloy systems // J. Electron. Mater. 2003. V. 32, P. 1265.
237. *Liu X. J.*, *Oikawa K.*, *Ohnuma I.*, *Kainuma R.*, *Ishida K.* The Use of Phase Diagrams and Thermodynamic Databases for Electronic Materials // JOM. 2003. V. 55. P. 53.
238. *Gaune-Escard M.*, *Bros J.-P.*, *Fouque Y.*, *Gaune P.*, *Hatem G.*, *Juhem P.* THERMOSALT, une banque de donnees thermodynamiques coherentes pour les melanges de sels fondees // Metaux. 1988. V. 64. P. 208.
239. *Zahra A. M.*, *Zahra C. Y.*, *Castanet R.*, *Jaroma-Weiland G.*, *Neuer G.* A data-bank for Thermophysical properties of light metals alloys // J. Therm. Anal. 1992. V. 38. P. 781.
240. *Östhols E.*, *Wanner H.* The NEA thermochemical data base project. OECD Nuclear Energy Agency / The NEA thermochemical data base project. OECD Nuclear Energy Agency. 2000. 21 p.
241. *Baba T.*, *Yamashita Y.*, *Nagashima A.* Function Sharing and Systematic Collaboration between a Networking Database System and Printed Media on Thermophysical Properties Data // J. Chem. and Eng. Data. 2009. V. 54. P. 2745.
242. *Villars P.*, *Berndt M.*, *Brandenburg K.*, *Cenzual K.*, *Daams J.*, *Hulliger F.*, *Massalski T.*, *Okamoto H.*, *Osaki K.*, *Prince A.*, *Putz H.*, *Iwata S.* The Pauling File, binaries edition // J. Alloys Compd. 2004. V. 367. P. 293.
243. *Kiselyova N.*, *Iwata S.*, *Dudarev V.*, *Prokoshev I.*, *Khorbenko V.*, *Zemskov V.* Integration Principles of Russian and Japanese Databases on Inorganic Materials // Information Technologies and Knowledge. 2008. V. 2. P. 366.
244. *Villars P.*, *Onodera N.*, *Iwata S.* The Linus Pauling file (LPF) and its application to materials design // J. Alloys Compd. 1998. V. 279. P. 1.

245. *Effenberg G.* The MSIT® workplace, access to materials chemistry data and knowledge / 6th International Scholl-Conference «Phase Diagrams in Materials Science». 14–20 October 2001. Technical Program & Abstracts. National Academy of Sciences of Ukraine. Kyiv. 2001.
246. *Pelton A. D.* Thermodynamic database development-modeling and phase diagram calculations in oxide systems // Rare Metals. 2006. V. 25. P. 473.
247. TAPP database, version 2.1, ESM software, Inc. Hamilton, USA. 1991–1994.
248. *Bergerhoff G.*, *Hundt R.*, *Sievers R.*, *Brown I. D.* The Inorganic Crystal Structure Data Base // J. Chem. Inf. and Comput. Sci. 1983. V. 23. P. 66.
249. *Bergerhoff G.* Data base for inorganic crystal structures // Comp. Phys. Communs. 1984. V. 33. P. 79.
250. *Fluck E.* Inorganic Crystal Structure Database (ICSD)and Standardized Data and Crystal Chemical Characterization of Inorganic StructureTypes (TYPIX) — Two Tools for InorganicChemists and Crystallographers // J. Res. NIST. 1996. V. 101. P. 217.
251. *Hellenbrandt M.* The inorganic crystal structure database (ICSD) — present and future // Crystallogr. Rev. 2004. V. 10. P. 17.
252. *Allmann R.*, *Hinek R.* The introduction of structure types into the Inorganic Crystal Structure Database ICSD // Acta Crystallogr. Sect. A. 2007. V. 63. P. 412.
253. *Mighell A. D.*, *Karen V. L.* NIST Crystallographic Databases for Research and Analysis // J. Res. NIST. 1996. V. 101. P. 273.
254. *Denley D. R.*, *Hart H. V.* RINGS: a new search/match database for identification by polycrystalline electron diffraction // J. Appl. Crystallogr. 2002. V. 35. P. 546–552.
255. *Byram S. K.*, *Campana C. F.*, *Fait J.*, *Sparks R. A.* Using NIST Crystal Data Within Siemens Software for Four-Circle and SMART CCD Diffractometers // J. Res. NIST. 1996. V. 101. P. 295.
256. *Wood G. H.*, *Rodgers J. R.*, *Gough S. R.* Operation of an international data center: Canadian Scientific Numeric Database Service // J. Chem. Inf. and Comput. Sci. 1993. V. 33. P. 31.
257. *White P. S.*, *Rodgers J. R.*, *Page Y. L.* CRYSTMET: a database of the structures and // Acta Crystallogr. Sect. B. 2002. V. 58. P. 343.
258. *Чичагов А. В.*, *Белоножко А. Б.*, *Лопатин А. Л.*, *Докина Т. Н.*, *Самохвалова О. Л.*, *Ушаковская Т. В.*, *Шилова З. В.* Информационно-вычислительная система по кристаллическим данным минералов (Минкрист) // Кристаллография. 1990. № 35. С. 610.
259. *Чичагов А. В.*, *Варламов Д. А.* Кристаллографическая и кристаллохимическая база данных для минералов и их структурных аналогов (WWW-МИНКРИСТ) / Теория, история, философия и практика минералогии: Материалы IV Международного минералогического семинара. Сыктывкар: Геопринт, 2006. С. 295.
260. *Kabekkodu S. N.*, *Faber J.*, *Fawcett T.* New Powder Diffraction File (PDF-4) in relational database format: advantages and data-mining capabilities // Acta Crystallogr. Sect. B. 2002. V. 58. P. 333.

261. *Allen F. H.*, *Davies J. E.*, *Galloy J. J.*, *Johnson O.*, *Kennard O.*, *Macrae C. F.*, *Mitchell E. M.*, *Mitchell G. F.*, *Smith M.*, *Watson D. G.* The development of versions 3 and 4 of the Cambridge structural database system // J. Chem. Inf. and Comput. Sci. 1991. V. 31. P. 187.
262. *Watson D. G.* The Cambridge Structural Database (CSD): Current Activities and Future Plans // J. Res .NIST. 1996. V. 101. P. 226.
263. *Allen F. H.* The Cambridge Structural Database: a quarter of a million crystal structures and rising // Acta Crystallogr. Sect. B. 2002. V. 58. P. 385.
264. *Dong Q.*, *Dewan A. K. R.*, *Marsh K. N.* DIPPR Project 882: Transport Properties and Related Thermodynamic Data for Binary Mixtures // Int. J. Thermophysics. 1999. V. 202. P. 237.
265. *Thomson G. H.*, *Larsen A. H.* DIPPR: Satisfying Industry Data Needs // J. Chem. and Eng. Data. 1996. V. 41. P. 930.
266. *Wilding W. V.*, *Rowley R. L.*, *Oscarson J. L.* DIPPR Project 801 evaluated process design data // Fluid Phase Equil. 1998. V. 150–151. P. 413.
267. *Киселева Н.*, *Мурат Д.*, *Столяренко А.*, *Дударев В.*, *Подбельский В.*, *Земсков В.* База данных по свойствам тройных неорганических соединений «Фазы» в сети Интернет // Информационные ресурсы России. 2006. № 4. С. 21.
268. *Киселева Н. Н.*, *Дударев В. А.*, *Земсков В. С.* Компьютерные информационные ресурсы неорганической химии и материаловедения // Успехи химии. 2010. Т. 79. № 2. С. 162–188.
269. *Дегтярев Ю. И.*, *Подбельский В. В.*, *Киселева Н. Н.*, *Петухов В. В.*, *Шеханова О. В.*, *Буш А. А.*, *Белокурова И. Н.*, *Пушко В. А.* База данных по свойствам кристаллов акустооптических, электрооптических и нелинейнооптических веществ, доступная из Internet // Изв. вузов. Материалы электронной техники. 1999. V. 3. P. 35.
270. *Xu Y.*, *Yamazaki M.*, *Wang H.*, *Yagi K.* Development of an Internet system for composite design and thermophysical property prediction // Mater. Trans. 2006. V. 47. P. 1882.
271. *Дударев В. А.* Международная интеграция баз данных по свойствам неорганических веществ / VIII российская ежегодная конференция молодых научных сотрудников и аспирантов «Физико-химия и технология неорганических материалов». Сборник материалов. М.: ИМЕТ РАН, 2011. С. 158–159.
272. *Dudarev V. A.*, *Kiselyova N. N.*, *Xu Y.*, *Yamazaki M.* Virtual integration of the Russian and Japanese databases on properties of inorganic substances and materials / MITS 2009. Symposium on Materials Database. National Institute for Materials Science (NIMS). Materials Database Station (MDBS). 2009. P. 37–48.
273. *Магарилл С. А.*, *Борисов С. В.*, *Подберезская Н. В.*, *Ипатова Е. Н.*, *Титов В. А.*, *Кузнецов Ф. А.* Кристаллические структуры неорганических веществ — база количественных данных. Принципы построения и опыт эксплуатации // Ж. структ. химии. 1995. № 36. С. 559.

274. *Девятых Г. Г.*, *Карпов Ю. А.*, *Осипова Л. И.* Выставка-коллекция веществ особой чистоты. М.: Наука, 2003.
275. *Asada Y.*, *Nakada E.*, *Yokokawa T.*, *Kurihara Y.*, *Yoshikawa A.* Database for materials design of multi-layered superconductors // J. Mater. Sci. Soc. Jap. 1988. V. 24. P. 199.
276. *Scott D. J.*, *Manos S.*, *Coveney P. V.*, *Rossiny J. C. H.*, *Fearn S.*, *Kilner J. A.*, *Pullar R. C.*, *Alford N.Mc N.*, *Axelsson A.-K.*, *Zhang Y.*, *Chen L.*, *Yang S.*, *Evans J. R. G.*, *Sebastian M. T.* Functional Ceramic Materials Database: An Online Resource for Materials Research // J. Chem. Inf. Model. 2008. V. 48. P. 449.
277. *Komornicki S.*, *Streiff R.* Relational data structure of the coating database from the Coatings & High Temperature Corrosion Data Bank // J. Phys. IV. 1993. V. 3. P. 1013.
278. *Ковач Э. А.*, *Лосев С. А.*, *Сергиевская А. Л.* Опыт создания автоматизированной системы научных исследований в области физико-химической газодинамики (система АВОГАДРО) // Металлы. 1993. № 4. С. 70.
279. *Кузнецова Л. А.*, *Пазюк Е. А.*, *Столяров А. В.* Банк данных РАДЭН. Радиационные и энергетические характеристики двухатомных молекул // Химия/МГУ. М.: МГУ, 1994. С. 53.
280. *Nebel A.*, *Tolle U.*, *Maass R.*, *Olbrich G.*, *Deplanque R.*, *Lister P.* The Integrated Gmelin Information System New developments in information processing // Anal. Chim. Acta. 1992. V. 265. P. 305.
281. *Lohr A.*, *Mez-Starck B.*, *Schirdewahn H.-G.*, *Watson D.-G.* MOGADOC (molecular gas-phase documentation) — an interactive computerized search/ retrieval system // J. Mol. Struct. 1983. V. 97. P. 57.
282. *Vogt J.* MOGADOC — A Bibliographic Numerical Resource for Gasphase Molecular Spectroscopy and Structure // J. Mol. Spectrosc. 1992. V. 155. P. 249.
283. *Talbot-Besnard S.*, *Rubinstein M.*, *Droniou C.* Les banques de donnees. Hydrogen-data // Ann.Chim.(Fr.). 1988. V. 13. P. 611.
284. *Дударев В. А.* Принципы интеграции БД по свойствам неорганических веществ и материалов / Теплофизические свойства веществ и материалов. Тезисы докладов XIIроссийской конференции по теплофизическим свойствам веществ. М.: Интерконтакт Наука, 2008. С. 186.
285. *Priven A. I.*, *Mazurin O. V.* Glass Property Databases: Their History, Present State, and Prospects for Further Development // Adv. Mater. Res. 2008. V. 39–40. P. 147.
286. *Mazurin O. V.* Glass properties: compilation, evaluation, and prediction // J. Non-Cryst. Solids. 2005. V. 351. P. 1103.
287. *Mazurin O. V.*, *Gankin Yu.* About testing the reliability of glass property data in binary systems // J. Non-Cryst. Solids. 2004. V. 342. P. 166.
288. *Мазурин О. В.*, *Гусаров В. В.* Будущее информационных технологий в материаловедении // Физика и химия стекла. 2002. № 28. С. 74.

289. *Saitou T.*, *Oguro H.*, *Fukami T.*, *Iseda T.* Constructing a Glass Material Database Using Java Language // Proceedings of the Japan Society of Information and Knowledge. May 1999. P. 63.
290. *Nagy M.*, *Over H. H.*, *Wollfart E.* XML related data exchange from the test machine to a web-enabled MAT-DB // Data Science J. 2005. V. 4. P. 151.
291. *Over H. H.*, *Wolfart E.*, *Dietz W.*, *Toth L.* Mat-DB: A Web-Enabled Materials Database to Support European R&D Projects and Network Activities // Adv. Eng. Mater. 2005. V. 7. P. 766.
292. *Janovec V.*, *Tomaszewski P. E.*, *Richterova L.*, *Fabry J.*, *Kluiber Z.* Inverse Database of Phase Transitions in Crystals with a Single Phase Transition // Ferroelectrics. 2004. V. 301. P. 169.
293. *Caracas R.* A database of incommensurate phases // J. Appl. Crystallogr. 2002. V. 35. P. 120.
294. *Tsuji H.*, *Yokoyama N.*, *Fujita M.*, *Kurihara Y.*, *Kano S.*, *Tachi Y.*, *Shimura K.*, *Nakajima R.*, *Iwata S.* Present status of Data-Free-Way (distributed database system for advanced nuclear materials) // J. Nucl. Mater. 1999. V. 271–272. P. 486.
295. *Kaji Y.*, *Miwa Y.*, *Tsukada T.*, *Tsuji H.*, *Nakajima H.* Status of JAERI material performance database (JMPD) and analysis of irradiation assisted stress corrosion caracking (IASCC) data // J. Nucl. Sci. and Techn. 2000. V. 37. P. 949.
296. *Kaji Y.*, *Tsuji H.*, *Fujita M.*, *Xu Y.*, *Yoshida K.*, *Mashiko S.*, *Shimura K.*, *Miyakawa S.*, *Ashino T.* Development of a knowledge based system linked to a materials database // Data Science J. 2004. V. 3. P. 88.
297. *Василенко Е.*, *Мещерякова Т.*, *Бацылев Ф.*, *Порысева Е.* Проблемно-ориентированная фактографическая база данных по нанокомпозитам // Информационные ресурсы России. 2009. № 4. С. 5.
298. *Дударев В. А.* Справочная система по информационным ресурсам неорганической химии с доступом из интернет / VII российская ежегодная конференция молодых научных сотрудников и аспирантов «Физико-химия и технология неорганических материалов». Сборник материалов. М.: Интерконтакт Наука, 2010. С. 131–132.
299. *Дударев В. А.* Подход к заполнению пропусков в обучающих выборках для компьютерного конструирования неорганических соединений // Вестник МИТХТ. 2014. Т. 9. № 1. С. 73–75.
300. *Столяренко А. В.* Интегрированная информационно-аналитическая система для прогнозирования свойств неорганических соединений: автореф. дис. … канд. техн. наук. М. 2008. 24 с.
301. *Литтл Р. Дж. А.*, *Рубин Д. Б.* Статистический анализ данных с пропусками. Пер. с англ. М.: Финансы и статистика, 1990. 336 с.
302. *Zloba E.*, *Yatskiv I.* Statistical methods of reproducing of missed data // Computer Modelling & New Technologies. 2002. V. 6. № 1. P. 51–61.
303. *Линник Ю. В.* Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений. 2-е изд. М.: 1962. 337 с.

304. *Li L.*, *Wolfel A.*, *Schonleber A.*, *Mondal S.*, *Schreurs A. M. M.*, *Kroon-Batenburg L. M. J.*, *van Smaalen S.* Modulated anharmonic ADPs are intrinsic to aperiodic crystals: a case study on incommensurate Rb_2ZnCl_4 // Acta Crystallogr. 2011. V. B67. № 3. P. 205–217.
305. *Sanctuary R.*, *Jundt D.*, *Baumert J.-C.*, *Gunter P.* Nonlinear optical properties of Rb_2ZnCl_4 in in-commensurate and ferroelectric phases // Phys. Rev. B. 1985. V. 32. № 3. P. 1649–1660.
306. *Белов К. П.*, *Третьяков Ю. Д.*, *Гордеев И. В. и др*. Магнитные полупроводники — халькогенидные шпинели. М.: МГУ, 1981. 279 с.
307. *Senko O.*, *Dokukin A.* Optimal Forecasting Based Convex Correcting Procedures / New Trends in Classification and Data Mining. Sofia: ITHEA. 2010. P. 62–72.
308. *Kiselyova N. N. Stolyarenko A. V.*, *Ryazanov V. V.*, *et al.* A system for computer-assisted design of inorganic compounds based on computer training // Pattern Recognition and Image Analysis. 2011. V. 21. № 1. P. 88–94.
309. http://www.w3.org/TR/cors
310. *Дударев В. А.*, *Воякин А. С.* SimpleCMS. Свидетельство о государственной регистрации программы ЭВМ № 2010615355 от 20.08.2010 г.
311. *Kiselyova N. N.*, *Dudarev V. A.*, *Zemskov V. S.* Computer information resources in inorganic chemistry and materials science // Russian Chem. Rev. 2010. V. 79. № 2. P. 145–166
312. *Киселева Н. Н.*, *Дударев В. А.* База данных «Информационные ресурсы неорганической химии и материаловедения» // Информационные технологии. 2010. № 12. С. 63–66.
313. *Дударев В. А.* База данных по информационным ресурсам в области неорганического материаловедения / VI российская ежегодная конференция молодых научных сотрудников и аспирантов. Сборник статей. М.: Интерконтакт наука, 2009. С. 127–129.

Приложение

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2010615355

SimpleCMS

Правообладатель(ли): ***Дударев Виктор Анатольевич (RU),***
Воякин Алексей Сергеевич (RU)

Автор(ы): ***Дударев Виктор Анатольевич,***
Воякин Алексей Сергеевич (RU)

Заявка № **2010613542**
Дата поступления **22 июня 2010 г.**
Зарегистрировано в Реестре программ для ЭВМ
20 августа 2010 г.

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Б.П. Симонов

РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2010615354

Data Entity Framework

Правообладатель(ли): ***Дударев Виктор Анатольевич (RU),***
Воякин Алексей Сергеевич (RU)

Автор(ы): ***Дударев Виктор Анатольевич,***
Воякин Алексей Сергеевич (RU)

Заявка № **2010613541**
Дата поступления **22 июня 2010 г.**
Зарегистрировано в Реестре программ для ЭВМ
20 августа 2010 г.

Руководитель Федеральной службы по интеллектуальной собственности, патентам и товарным знакам

Б.П. Симонов

МИНИСТЕРСТВО СВЯЗИ И МАССОВЫХ КОММУНИКАЦИЙ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ИНФОРМАЦИОННЫМ ТЕХНОЛОГИЯМ

ФГУП Научно-технический центр "Информрегистр"

ГОСУДАРСТВЕННЫЙ РЕГИСТР БАЗ ДАННЫХ

РЕГИСТРАЦИОННОЕ СВИДЕТЕЛЬСТВО

№ 12242

от "24" февраля 2009 г.

Настоящее свидетельство выдано организации:

Учреждение Российской академии наук Институт металлургии и материаловедения им. А.А. Байкова РАН

в том, что представленная в Государственный регистр база данных

База данных по ширине запрещенной зоны неорганических веществ "Bandgap"

зарегистрирована за **№ 0220913091**

М.П. Научно-технический центр "Информрегистр" Москва

Директор ФГУП НТЦ "Информрегистр"

Е.И. Козлова

"24" февраля 2009 г.