Государственное образовательное учреждение высшего профессионального образования "Московская государственная академия тонкой химической технологии им. М.В. Ломоносова"

На правах рукописи

Дударев Виктор Анатольевич

ИНТЕГРИРОВАННАЯ СИСТЕМА БАЗ ДАННЫХ ДЛЯ ИНФОРМАЦИОННОЙ ПОДДЕРЖКИ ПРИНИЯТИЯ РЕШЕНИЙ ПРИ ПРОГНОЗИРОВАНИИ СВОЙСТВ НЕОРГАНИЧЕСКИХ ВЕЩЕСТВ

05.13.01 – Системный анализ, управление и обработка информации (химическая технология)

Диссертация на соискание ученой степени кандидата технических наук

Научный руководитель
д.т.н., профессор Корнюшко В.Ф.

Москва – 2006

Содержание

ВВЕДЕНИЕ 3
ГЛАВА 1. СИСТЕМНЫЙ ПОДХОД К РАЗРАБОТКЕ ИНТЕГРИРОВАННОЙ СИСТЕМЫ ПО СВОЙСТВАМ ВЕЩЕСТВ ДЛЯ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ 7
1.1. Технологии интеграции гетерогенных информационных систем 9
1.2. БД по свойствам веществ для электронной техники, созданные в мире 31
1.3. Выбор платформы для разработки распределенной информационной системы 39
Краткие выводы 50
ГЛАВА 2. МЕТОДИКА ПОСТРОЕНИЯ ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ПО СВОЙСТВАМ ВЕЩЕСТВ ДЛЯ ЭЛЕКТРОНИКИ 51
2.1. Информационные системы по свойствам неорганических веществ ИМЕТ РАН 51
2.2. Расчетные подсистемы информационных систем по свойствам неорганических веществ 60
2.3. Архитектура современных информационных систем по свойствам веществ 64
2.4. Комплексный подход к интеграции информационных систем 67
2.5. Интеграция распределенных гетерогенных Web-приложений информационных систем 72
2.6. Интеграция гетерогенных источников данных информационных систем 80
Краткие выводы 94
ГЛАВА 3. РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ 95
3.1. Реализация интеграции гетерогенных Web-приложений информационных систем 95
3.2. Реализация интеграции гетерогенных источников данных информационных систем 110
Краткие выводы 127
ГЛАВА 4. ПРИМЕНЕНИЕ ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ДЛЯ ПОИСКА ЗАКОНОМЕРНОСТЕЙ И КОМПЬЮТЕРНОГО КОНСТРУИРОВАНИЯ НОВЫХ СОЕДИНЕНИЙ 129
4.1. Этапы компьютерного конструирования новых соединений 130
4.2. Перспективные полупроводники 133
4.3. Перспективные диэлектрики 137
Краткие выводы 139
ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ 140
ЛИТЕРАТУРА 141
ПРИЛОЖЕНИЕ 151

ВВЕДЕНИЕ

Обеспечение химиков-технологов достоверной информацией о свойствах и технологиях получения современных веществ является необходимым условием развития современной промышленности. На современном этапе качественная информационная поддержка специалистов невозможна без использования специализированных баз данных (БД). Разработка информационных систем (ИС) по свойствам веществ и процессам их получения ведется во всех промышленно развитых странах [1]. Многие ИС доступны в режиме удаленного доступа с использованием телекоммуникационных сетей [2, 3]. Наиболее мощные информационные системы, основанные на современных СУБД, предлагают NIST (National Institute of Standards and Technology – Национальный институт стандартов и технологий, США) и STN (The Scientific and Technical Information Network – Международная сеть научно-технической информации) [4-18]. Как правило, БД по свойствам веществ разрабатываются в разных организациях и даже в разных странах. Полная интеграция таких систем невозможна из-за разного уровня качества данных, хранящихся в разных БД ИС. Обычно она связана и с организационными трудностями, т.к. большинство ИС используются в коммерческих целях или являются открытыми для доступа пользователей только определенных стран или организаций.

В последние годы наблюдается тенденция к кооперации в разработке ИС и к интеграции уже созданных ИС, как на национальном, так и на международном уровне. Актуальность решения этой задачи вызвана стремлением устранить необоснованное дублирование работ и уменьшить затраты на разработку и поддержку ИС. Кроме того, интеграция информации, содержащейся в ИС по свойствам веществ и технологиям их получения, позволяет применять методы компьютерного 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.

Целью работы является информационная поддержка принятия решений при прогнозировании свойств веществ на основе интеграции разнородных баз данных по свойствам веществ и технологиям их получения.

Для достижения цели работы были поставлены следующие задачи:

- провести анализ современных технологий интеграции разнородных информационных систем;
- осуществить выбор программной платформы для построения интегрированной информационной системы;
- проанализировать и систематизировать архитектуру современных информационных систем по свойствам веществ и технологиям их получения;
- разработать методику построения интегрированной информационной системы с учетом возможности ее использования конечными пользователями и системами поддержки принятия решений;
- разработать структуры данных для применения в интегрированной информационной системе по свойствам веществ;
- разработать интегрированную информационную систему в виде программного комплекса;
- разработать программное обеспечение баз данных по свойствам акустооптических, электрооптических и нелинейнооптических веществ “Кристалл” и по ширине запрещенной зоны неорганических веществ “BandGap”;
- применить созданную интегрированную информационную систему для прогнозирования свойств веществ, перспективных для использования в современной промышленности.

Для достижения этой цели было необходимо найти решение проблем интеграции информационных систем, удовлетворяющее следующим условиям. Решение должно быть:

- масштабируемым, т.е. обеспечивать возможность поэтапного добавления существующих информационных систем;
- достаточно простым для реализации, чтобы на основе предложенной методики любой участник мог самостоятельно разработать программные модули для включения своей информационной системы в интегрированную систему;
- гибким, чтобы учитывать различия в данных и информационных структурах ИС разных организаций;
- мощным, чтобы обеспечить сложные механизмы извлечения и манипулирования данными.

Научная новизна

- предложен комплексный подход к интеграции ИС, как на уровне пользовательских интерфейсов, так и на уровне источников данных;
- на основе теории множеств дано определение релевантной информации в контексте интегрированной ИС по свойствам неорганических веществ;
- разработаны схемы данных и алгоритмы разрешения конфликтов гетерогенности для интегрированной ИС по свойствам веществ и технологиям их получения;
- разработана методика применения интегрированной ИС в программном комплексе компьютерного конструирования химических соединений для прогнозирования свойств веществ.

Практическая значимость

Разработан и внедрен в Институте металлургии и материаловедения им. А.А. Байкова РАН (ИМЕТ РАН) программный комплекс, реализующий интегрированную ИС, объединяющий информационные системы по свойствам веществ и технологиям их получения. При помощи этого программного комплекса выполнена интеграция информационных систем, разработанных ИМЕТ РАН совместно с другими организациями России: БД по свойствам неорганических соединений “Фазы”, БД по фазовым диаграммам полупроводниковых систем “Диаграмма”, БД по свойствам акустооптических, электрооптических и нелинейнооптических веществ “Кристалл”, БД по ширине запрещенной зоны неорганических веществ “BandGap”, БД по свойствам химических элементов “Элементы” и информационной системы по свойствам полупроводникового кремния и процессам его получения и обработки “Кремний”. Полученный информационный комплекс не только позволяет конечным пользователям получать доступ ко всей информации и расчетным подсистемам в рамках интегрированной ИС, но и использовать ИС в качестве источника информации для программ компьютерного конструирования соединений и СППР.

Применение интегрированной ИС позволяет сократить время, затрачиваемое на поиск полной информации по свойствам и технологиям получения веществ.

Полученный программный комплекс используется в учебном процессе кафедры "Материалы микро-, опто- и наноэлектроники" МИТХТ при чтении курсов "Технология полупроводниковых материалов" и "Моделирование процессов полупроводниковой технологии".

Методы исследования

Структуризация и формализация предметной области выполнена на основе методов структурного системного анализа. Для определения релевантной информации в контексте интегрированной ИС и построения модели понятий предметной области использован математический аппарат теории множеств. При разработке интегрированной ИС использованы теория построения БД и Web-технологии. Для иллюстрации использования ИС в интеллектуальных системах использованы методы индуктивного вывода и компьютерного конструирования неорганических соединений, основанные на обучении ЭВМ распознаванию образов.

Структура диссертации

Диссертация состоит из введения, 4 глав и заключения.

В первой главе рассматриваются наиболее значимые БД по свойствам неорганических веществ, созданные в мире, современные технологии интеграции ИС и анализируются программные платформы для построения интегрированной ИС. Во второй главе на основе структурного анализа архитектуры ИС предлагается комплексный подход к интеграции ИС и разрабатывается методика интеграции, как на уровне интерфейсов, так и на уровне источников информации ИС. В третьей главе описана реализация интегрированной ИС в виде программного комплекса. В четвертой главе представлены результаты применения интегрированной ИС в качестве источника данных для программ компьютерного конструирования неорганических соединений.

В заключении сформулированы основные результаты, полученные в диссертационной работе.

Исследования, представленные в настоящей диссертационной работе, были поддержаны российскими фондами и организациями: РФФИ (гранты №04-07-90086, 06-07-89120 и 05-03-39009) и Правительством Москвы (гранты №3-4 и 1.2.1 Программы «Инфраструктура и адресная поддержка науки»).

ГЛАВА 1. СИСТЕМНЫЙ ПОДХОД К РАЗРАБОТКЕ ИНТЕГРИРОВАННОЙ СИСТЕМЫ ПО СВОЙСТВАМ ВЕЩЕСТВ ДЛЯ СИСТЕМ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ

Существующие определения термина “система поддержки принятия решений” (СППР), как правило, основаны на описании целей и функций этой системы [19]. 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 [20] (рис. 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 [21]. 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). 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 четвертой главе настоящей работы.

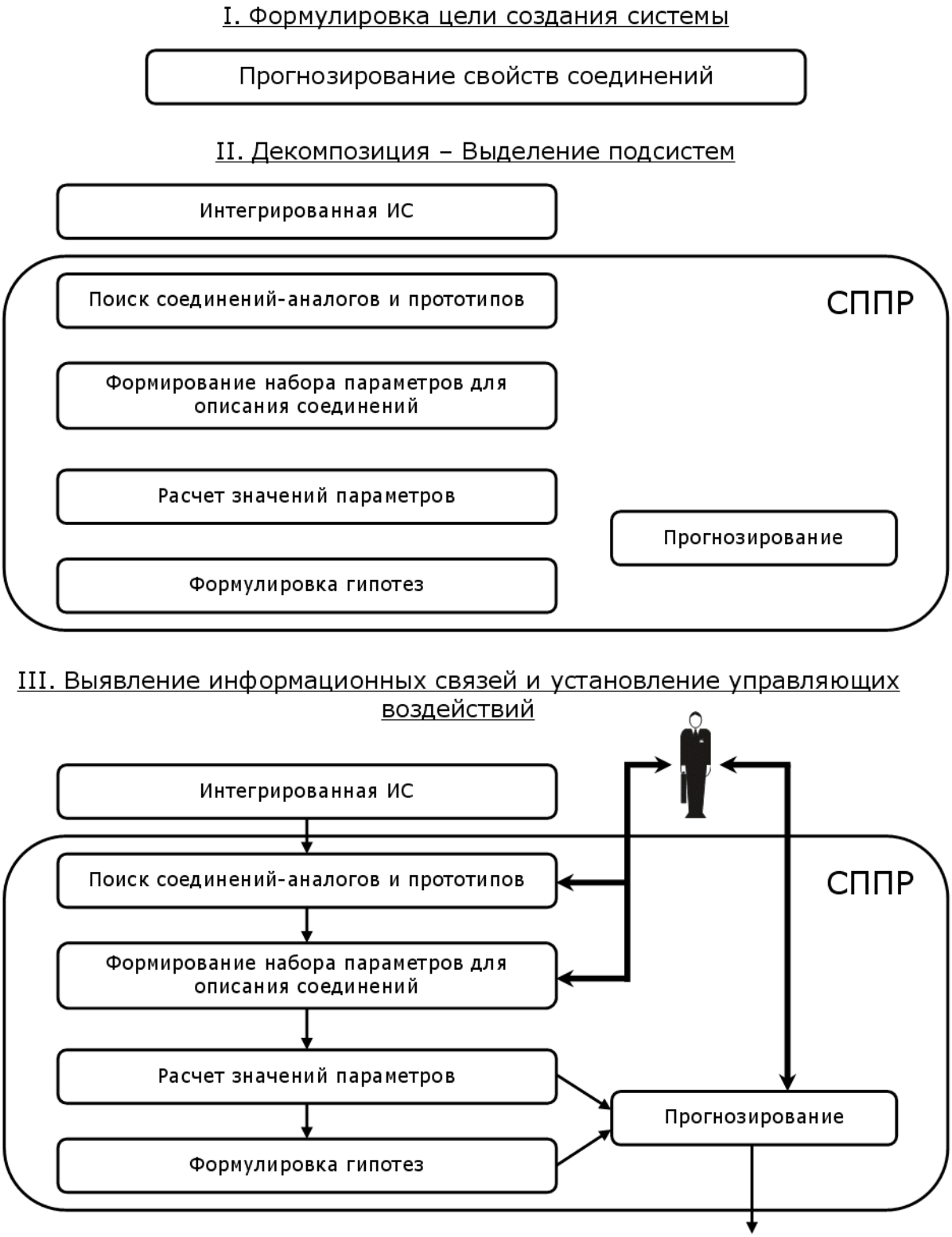

Рис. 1.1. Системный подход к разработке структурной схемы СППР.

Целью настоящей работы является интеграция баз данных по свойствам веществ и технологий их получения. Для ее достижения необходимо выполнить обзор современных технологий интеграции ИС, БД по свойствам веществ и технологий их получения и современных программных платформ для построения интегрированной ИС.

1.1. Технологии интеграции гетерогенных информационных систем

1.1.1. Актуальность интеграции

В настоящей диссертационной 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 и химических соединений. 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 – 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 диссертационной 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, в 2002 году затраты на интеграцию информационных систем и оценку качества данных по всему миру составили порядка одного триллиона долларов США [22]. Следует также отметить, что по данным Forrester, 33% всех компаний в сфере ИТ занимаются интеграцией информационных систем [23]. А обзор за февраль 2002 года, проведенный CIO Magazine, показывает, что самой высокоприоритетной статьей расходов многие ИТ-компании считают построение интегрированных систем (рис. 1.2).

Рис. 1.2. Наибольшие расходы ИТ-компаний по сферам деятельности.

1.1.2. 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 данных, 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 [24]. 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.

1.1.3. Подходы к интеграции

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 технологий 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 подходов к 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 технологий 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 настоящее время выделяют три подхода к 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).

Принципы интеграции, заложенные в этих подходах, используются для решения широкого круга задач: от интеграции в режиме реального времени до пакетной интеграции, и от интеграции данных до интеграции приложений. На рис. 1.3 показано положение названных технологий 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 являются технологии, основанные на подходе EAI. Следует отметить, что ни один из существующих на сегодняшний день подходов к интеграции не способен решить все проблемы, возникающие при объединении ИС.

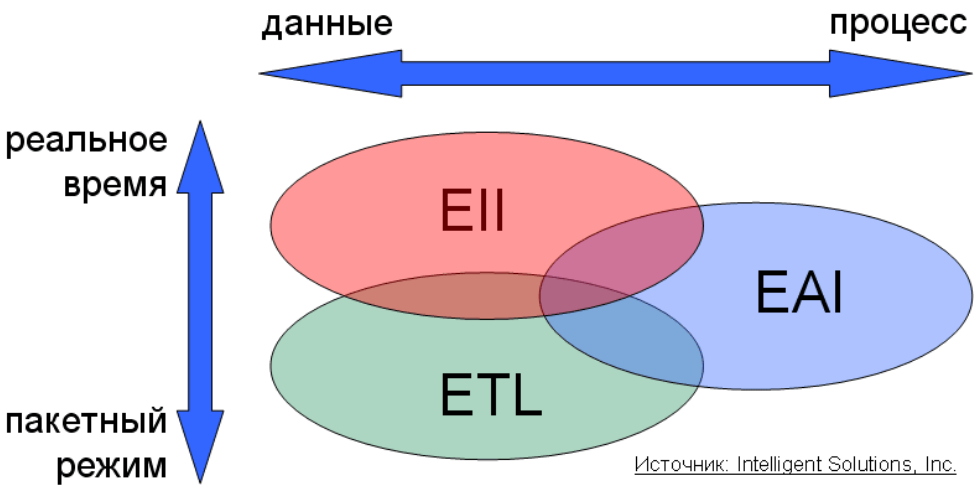

Рис. 1.3. Современные подходы к интеграции [25].

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 подходов, и в каких случаях тот или иной подход должен использоваться. Необходимо четко представлять возможности каждого подхода и определить класс задач, для решения которых подходит каждый из подходов. Для понимания различий в назначении интеграционных подходов необходимо привести соответствующие определения, учитывающие их назначения [25]:

- **EAI** – это интеграционный подход, с помощью которого организация добивается централизации и оптимизации интеграции корпоративных приложений, обычно

используя те или иные формы технологии оперативной доставки информации (push technology), которая управляется внешними событиями (event-driven);

- **ETL** – это интеграционный подход, 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. ETL-технология ориентирована на разные базы данных, например, хранилища данных, витрины или операционные склады данных;
- **EII** – это подход к 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. Подход EII ориентирован на конкретных сотрудников, которые получают информацию через специальную инструментальную панель или отчет.

Для более полного понимания этих подходов 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. 1.4 показано, как каждый из этих подходов может быть использован наилучшим образом. Подход EAI интегрирует транзакции двух или более приложений, подход ETL интегрирует данные операционных систем и компонентов поддержки принятия решений, а подход EII осуществляет виртуальную интеграцию данных из различных источников.

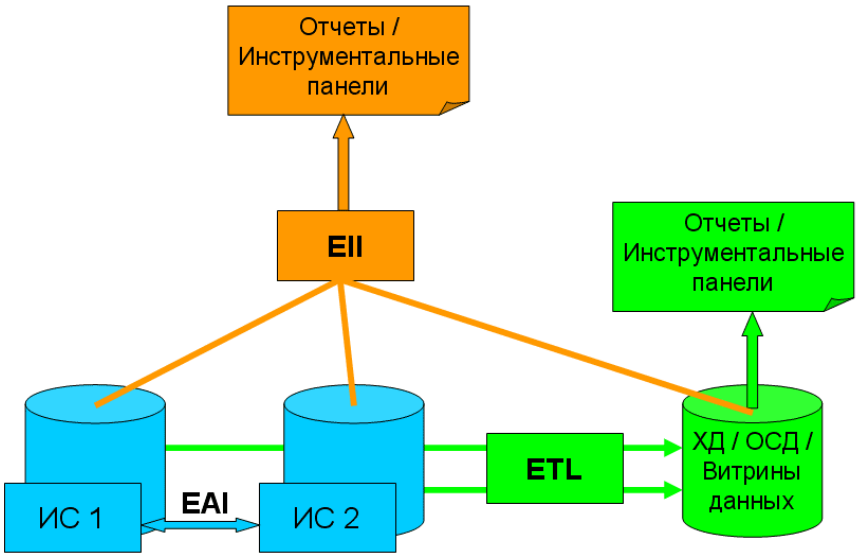

Рис. 1.4. Способы интеграции двух информационных систем с использованием подходов интеграции EII, ETL, EAI [25].

Проанализируем информационные потоки при разных методах интеграции ИС по свойствам веществ и технологии их производства, учитывая связь интегрированной ИС с системами поддержки принятия решений (СППР) при исследовании и производстве химических соединений для электронной техники (рис. 1.5).

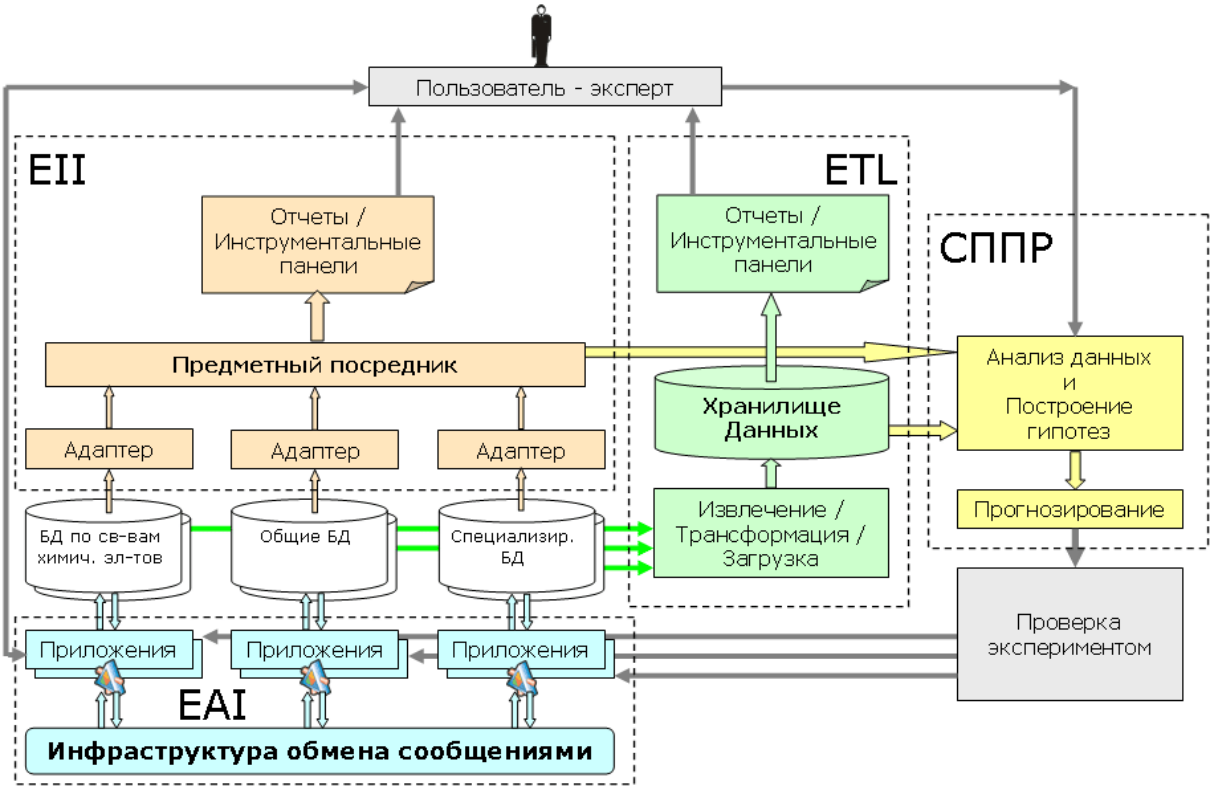

Рис. 1.5. Анализ информационных потоков при использовании разных подходов к интеграции и место интегрированной ИС в СППР.

Рассмотрим более подробно сценарии использования каждого из указанных подходов к интеграции. Подход 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). Подход EAI применяется также в ситуации, когда необходимо, чтобы изменения, внесенные в одну ИС (обычно это небольшой набор записей), были отражены во всех других ИС. Этот подход используется при решении задач фиксации изменений и их переноса в соответствующие ИС и часто применяется, например, в банковской сфере.

Интеграционный подход 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 подход также используется для интеграции ключевых справочных данных. ETL-подход 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).

Подход 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 подходы, нужно также знать и проблемы, которые им присущи. Во-первых, внедрение этих подходов к интеграции требует от ИТ-персонала глубокого понимания тех требований, которые предъявляются к данным для принятия как тактических, так и стратегических решений. Применительно к подходу 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 подходов 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 подходов к 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 подходов к интеграции очевидно, что в качестве такого интеграционного подхода целесообразно использовать EII, который позволяет в

режиме реального времени извлекать требуемые данные из виртуально интегрированной информационной системы.

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 этой технологии на рынке были созданы несколько EII-продуктов и накоплен значительный опыт применения [26].

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 технологии 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 [26]:

- 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 информационной системы 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. EII-технологии используют 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 специализированного 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: (a) максимизации параллельности обработки запросов и (b) 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 до сих пор не выпустила окончательную версию спецификации XQuery 1.0 (текущая версия спецификации – XQuery 1.0: An XML Query Language, Candidate Recommendation 3rd November 2005 [27]). Консорциумом W3C стандартизованы пока только XPath 1.0 и XSLT 1.0, 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 технологии 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 зрелой технологии хранилищ данных. Последняя изначально рассматривалась как технология получения ответов на запросы в режиме реального времени. Тем не менее, с увеличением мощности современных компьютеров пока еще более медленные технологии 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 технологии 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 [26] в 2005 году ожидается, что доходы

индустрии EII составят, по меньшей мере, 0,5 млрд. долларов США. Вместе с тем очевидно, что те продукты, которые используются в настоящее время, уже в скором будущем претерпят значительные изменения, что позволит реализовать полный потенциал EII-подхода к интеграции.

У находящейся фактически в зачаточном состоянии EII существует уже достаточно зрелая альтернатива в лице *технологии хранилищ данных (ETL)*. Еще в начале 90-х годов У. Инмон определил хранилища данных (Data Warehouse) [28] как “предметно-ориентированный, интегрированный, неизменчивый, поддерживающий хронологию набор данных, организованный для целей поддержки управления”.

За последние пятнадцать лет технологии 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 [29].

Потребность в снижении цен на готовые ИТ-решения и необходимость обработки запросов в режиме реального времени приводит к разработке и апробации технологий 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-технологии. 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 даже несмотря на то, что EII находится лишь в начале своего развития, она должна заменить технологии хранилищ данных. Следует отметить, что это мнение ошибочно, и привести два аргумента против него:

- Для того чтобы быть жизнеспособными, EII-технологии 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 [30].
- Даже если EII-технологии будут полностью доработаны, они все же не заменят технологий 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 [29].

Таким образом, предположение о том, что EII заменит хранилища данных, поверхностно и неверно. Технологии хранилищ данных будут продолжать использоваться для построения больших систем и решения сложных аналитических задач. Подход 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 [29]:

- **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 [29]:

- **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.

- **Виртуализируйте 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.
- **Виртуализируйте 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 технологиями хранилищ данных (ETL) является то, что EII – технология интеграции данных “по требованию” (или технология “вытягивания” данных из источников в режиме реального времени), в то время как ETL – технология, при которой разрозненные данные помещаются в единое информационное хранилище данных заранее и, фактически, полностью готовы к запросам [31]. 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 технологии 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 технологии интеграции – это проблема постановки задачи. Каждая из рассмотренных выше технологий (EII, ETL, EAI) эффективны для решения различных классов задач, поэтому только в редких случаях могут заменять друг друга.

1.1.4. 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.

В настоящее время существует 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. 1.6). 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.

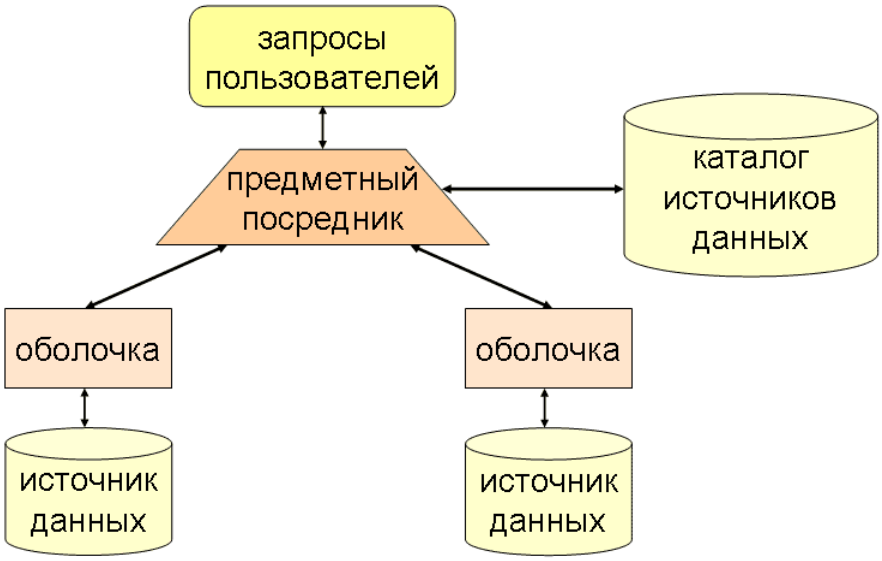

Рис. 1.6. Схема интеграции источников данных на основе схемы с участием предметного посредника (медиаторная схема).

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. 1.7 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.

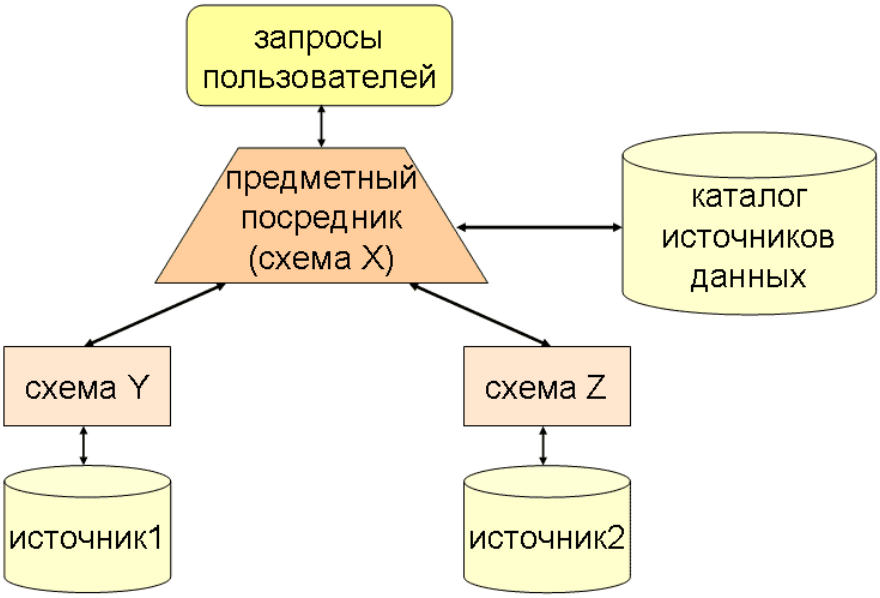

Рис. 1.7. Интеграция на основе принципа Global-As-View.

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 [32], MOMIS[33] и IBIS [34].

При подходе Local-As-View [35-37] все конструкции локальной схемы определены как представления глобальной схемы. Каждый источник данных описывается одним или несколькими представлениями согласно общей схеме предметного посредника (рис. 1.8). 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.

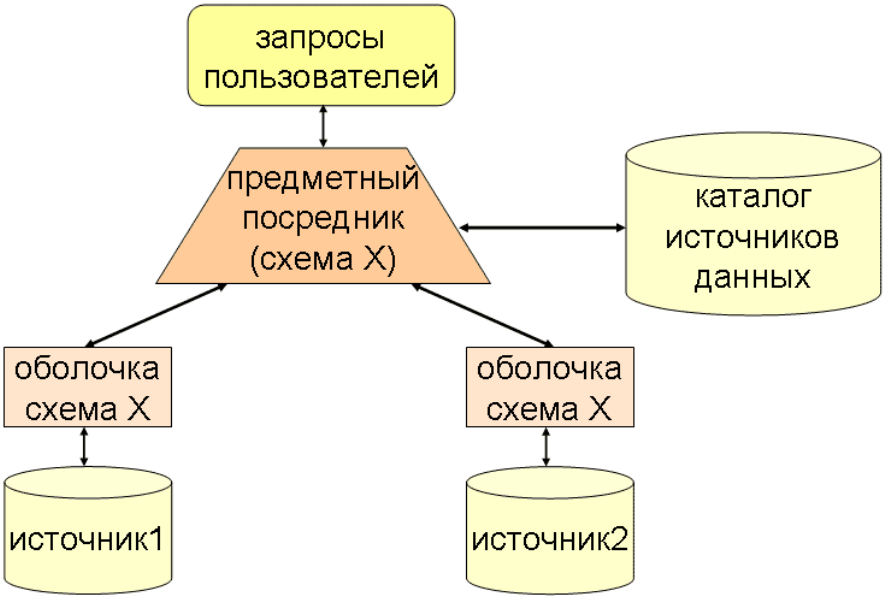

Рис. 1.8. Интеграция на основе принципа Local-As-View.

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 [38].

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 [39]. 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 [40-

43]. 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.

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.

1.1.5. 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 [44]:

- **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”, “CaTiO3”, “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, предъявляемыми 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.

1.2. БД по свойствам веществ для электронной техники, созданные в мире

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 [4-18, 45-101]. 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 диссертационной работе рассмотрены базы данных по свойствам неорганических веществ, и в первую очередь – БД по свойствам веществ, используемых в электронике.

Информацию о веществах для электроники можно разбить на три группы: (1) данные о веществах для использования в качестве <u>активных</u> компонентов (полупроводниковых устройств, магнитной памяти, пьезоэлектрических преобразователей, фильтров и гетеродинов, пиро- и сегнетоэлектрических, лазерных, сверхпроводящих, нелинейнооптических, акустооптических, электрооптических устройств и т.д.), (2) информация о веществах для применения в качестве <u>пассивных</u> компонентов (резисторов, трансформаторов, проводников, оптических волокон, печатных плат и т.д.) и (3) данные о <u>вспомогательных</u> 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. 1.1 дан перечень некоторых БД по свойствам неорганических веществ, которые содержат информацию о веществах, используемых в электронике [1].

Таблица 1.1. Базы данных по свойствам неорганических веществ, используемых в электронике.

Название БД	**Организация, страна**	**Предметная область**	**Ссылка**
Inorganic Crystal Structure Database (ICSD)	FIZ Karlsruhe, ФРГ, National Institute of Standards and Technology (NIST) (Gaithersburg), США	Кристаллические структуры неорганических соединений	4
CRYSTMET	National Research Council (NRC), Канада	Кристаллические структуры интерметаллических соединений	5
NEWCRYST	FIZ Karlsruhe, Oldenbourgh Verlag, Германия	Кристаллическая структура органических и неорганических веществ	12
Crystal Data (CD)	NIST, США	Кристаллические структуры неорганических соединений	6
Electron Diffraction Database (EDD)	NIST, Sandia National Laboratory (Albuquerque), США	Рентгенограммы порошка, кристаллографические данные, сведения по дифракции электронов	7, 8
БД по структурным фазовым переходам	Институт низких температур и структурных исследований Польской Академии наук (Вроцлав), Польша	Данные о структурных фазовых переходах в кристаллах	8
Pauling File on Binary Systems	Japan Science and Technology Corporation (Tokyo), Япония, Material Phases Data System (Vitznau), Швейцария	Фазовые диаграммы двойных систем, кристаллические структуры и физические свойства двойных неорганических соединений	45
MSI-Phase Diagram Center	Materials Science Int. Services, GmbH, ФРГ	Фазовые диаграммы металлических систем	56
Powder Diffraction File (PDF)	International Center for Diffraction Data (ICDD), США	Порошковые дифрактограммы неорганических и органических веществ	57

Название БД	**Организация, страна**	**Предметная область**	**Ссылка**
ICSDB (Incommensurate Structures Database)	Universidad del Pais Vasco (Bilbao), Испания	Структура несоразмерных фаз	58
БД по свойствам материалов электронной техники (СМЭТ)	ИНХ СО РАН (Новосибирск), Россия	Физические, термодинамические и кристаллоструктурные свойства полупроводниковых материалов	59-62
Cambridge Structural Database (CSD)	Cambridge University, Англия	Кристаллическая структура комплексных соединений с органическими лигандами	63
ИВТАНТЕРМО	ОИВТ РАН (Москва), Россия	Теплофизические (рекомендованные) данные для неорганических веществ	64
БД по физико-химическим свойствам высокочистых веществ	ИХВВ РАН, (Нижний Новгород), Россия	Физико-химические свойства высокочистых веществ, определяющие эффективность процессов очистки; коэффициент распределения примеси при двухфазных равновесиях	60, 65
THERMODATA	Информационный центр по химической термодинамике минералов (Гренобль), Ун-т в Сан Мартен д'Эр, Лаборатория термодинамики и физической химии металлов, Франция	Библиография по термодинамическим свойствам неорганических соединений и сплавов и по фазовым диаграммам	56, 66
THERMALLOY	Информационный центр по химической термодинамике минералов (Гренобль), Ун-т в Сан Мартен д'Эр, Франция	Термодинамические свойства неорганических соединений и сплавов	56

Название БД	**Организация, страна**	**Предметная область**	**Ссылка**
THERMCOMP	Информационный центр по химической термодинамике минералов (Гренобль), Ун-т в Сан Мартен д'Эр, Франция	Термохимические свойства неорганических соединений	56
MTDATA	National Physical Laboratory (Teddington, Middlesex), Англия	Термодинамические свойства неорганических соединений и сплавов, фазовые диаграммы	9
NIST Phase Equilibria Diagrams DB	NIST (Gaithersburg), США	Фазовые диаграммы неорганических систем	10
TRCTHERMO	Thermodynamic Research Center of Texas A & M University, США	Термодинамические свойства веществ	67
NISTTHERMO	NIST, США	Термохимические свойства неорганических и малых органических молекул	68
NIST Thermodynamic and Transport Properties of Pure Fluids DB	NIST (Gaithersburg), США	Теплофизические свойства чистых неорганических и органических жидкостей и газов	11
JICST TH	Japan Information Center of Science and Technology (JICST), Япония	Термофизические и химические свойства неорганических веществ	69
JANAF	NIST (Gaithersburg), США	Термофизические и химические свойства неорганических веществ и органических веществ с одним-двумя атомами углерода	68

Название БД	**Организация, страна**	**Предметная область**	**Ссылка**
DETHERM	DECHEMA/FIZ CHEMIE, ФРГ	Термофизические свойства, фазовые равновесия, транспортные и электрические свойства неорганических соединений и смесей; свойства поверхности и растворов электролитов	70
DIPPR	AIChE Design Institute for Physical Prop erty Data, США	Свойства промышленных веществ, зависящие от температуры	71, 72
БД по химической кинетике	NIST, США	Газофазные кинетические данные	73
БД по термофизическим свойствам газов	NIST, США	Термофизические свойства газов, используемых в полупроводниковой промышленности	74
TEPIAC- TPRC	Purdue University (Center for Information and Numerical Data Analysis and Synthesis - CINDAS), (Вест-Лафайетт, шт. Индиана), США	Теплофизические и электронные свойства неорганических веществ и материалов	75, 76
АСТРА	Московский государственный технический университет (Москва), Россия	Теплофизические (рекомендованные) данные для неорганических веществ	77
SGTE	Scientific Group Thermodata Europe, страны ЕЭС	Теплофизические данные для неорганических веществ	78
SGTE Unary DB	Scientific Group Thermodata Europe, страны ЕЭС	Теплофизические данные для химических элементов	79
THERMO-CALC	Королевский технологический ин-т, Швеция	Термохимические свойства неорганических соединений и сплавов	80

Название БД	**Организация, страна**	**Предметная область**	**Ссылка**
Gmelin-Online Datensystem	Gmelin Institute for Inorganic Chemistry, ФРГ	Физико-химические свойства неорганических и металлоорганических соединений	81
Atomic Energy Level Data Center	NIST (Optical Physics Division), США	Атомные спектры неорганических веществ	12
MOGADOC	University of Ulm, ФРГ	Библиография по структурным, электрическим и магнитным свойствам, электронным и микроволновым спектрам молекул неорганических, элементоорганических и органических веществ, находящихся в газовой фазе	82
Electronic Materials Information Service (EMIS)	Institute of Electrical Engineers, Michael Faraday House, Англия	Библиография по полупроводниковым и другим материалам электронной техники	56
Optical Materials Properties	NIST, Sandia National Laboratory (Albuquerque), США	Оптические свойства пластмасс, покрытий и стекол при 300-2500 нм	56
NEL-APPES	National Engineering Laboratory (Glasgow), Англия	Физические свойства неорганических веществ	83
SUPCONA	Национальный исследовательский Институт металлов (NRIM) (Токио), Япония	Свойства сверхпроводников	84
NIST/NRIM High Temperature Superconductors Database	NIST (Gaithersburg), США, NRIM, Япония	Свойства сверхпроводников	13
БД по сверхпроводникам	Токийский университет, Япония	Свойства сверхпроводников	85
ASMDATA	ASM International (Materials Park, Ohio), США	Свойства композитов, сталей, металлов, цветных сплавов и пластиков	86

Название БД	Организация, страна	Предметная область	Ссылка
COPPERDATA	Copper Development Association, Inc. (New York), США	Физические, механические, электрические, коррозионные и термические свойства меди и ее сплавов	87
Rare Earth Information Center	Iowa State University, Energy Resource and Development Administration, США	Свойства р.з.м.	88
CIAC	Ин-т прикладной химии (CIAC), Китай	Физико-химические свойства р.з.м. и параметры процессов экстракции	89
MDF	Materials Information ASM International Materials Park (Ohio), США, The Institute of Metals (London), Англия	Механические и физические свойства железа, цветных металлов и их сплавов	90
БД по аморфным материалам	Institute for Materials Research, Япония	Свойства аморфных материалов	91
БД по свойствам люминофоров	НПО "Люминофор" (Ставрополь), Россия	Свойства люминофоров	89
NISTCERAM	NIST (Gas Research Institute (GRI)), США	Термические, механические и коррозионные свойства карбида и нитрида кремния	56, 14
Microelectronics Packaging Materials Database (MPMD)	Purdue University (CINDAS), США	Физические и физико-химические свойства материалов, применяемых в микроэлектронике	76
Surface Structure Database (SSD)	NIST, США	Структура поверхности неорганических и органических веществ	15
ЭПИДИФ	ОИВТ РАН (Москва), Россия	Коэффициенты взаимной диффузии компонентов газофазной эпитаксии полупроводниковых материалов	93

Название БД	Организация, страна	Предметная область	Ссылка
БД по процессам получения эпитаксиальных гетероструктур полупроводнико-вых материалов методом жидкофазной эпитаксии	МИТХТ (Москва), Россия	Информация о параметрах процессов получения эпитаксиальных гетероструктур полупроводниковых материалов методом жидкофазной эпитаксии	60, 94

На рис. 1.9 дано распределение 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.

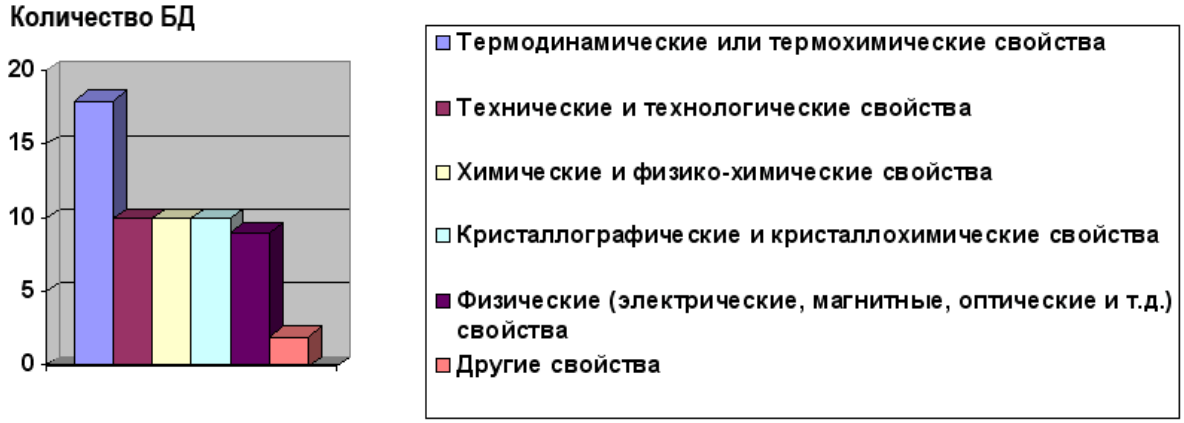

Рис. 1.9. Распределение БД по свойствам неорганических веществ по тематике.

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 [4-18].

Следует отметить, что только несколько баз данных, указанных в таблице 1.1, созданы для информационного обеспечения специалистов, разрабатывающих и использующих вещества, используемые в электронной технике: СМЭТ [59-62], БД по сверхпроводникам, разработанная в Токийском университете [85], SUPCONA [84], NIST/NRIM High Temperature Superconductors Database [6], NISTCERAM [7], ЭПИДИФ [93], БД, разработанная в МИТХТ [60, 94], Microelectronics Packaging Materials Database, предлагаемая информационным центром CINDAS при Purdue University [76].

Базы данных по свойствам веществ для электроники ИМЕТ РАН [47, 51, 53]: 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.

1.3. Выбор платформы для разработки распределенной информационной системы

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 [102].

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).

1.3.1. Производительность

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

- Web-сервер Internet Information Server (IIS) на платформе Microsoft Windows Server 2003;
- Web-сервер Apache на базе высокопроизводительных коммерческих Unix-платформ.

Для сравнения используем результаты отчета “Microsoft Windows Server 2003 with Internet Information Services (IIS) 6.0 vs. Linux Competitive Web Server Performance Comparison”, опубликованного Veritest (www.veritest.com) в апреле 2003 года [103]. В этих тестах независимой компанией сравнивалась пиковая производительность по числу обрабатываемых запросов в секунду, показываемая указанными выше Web-серверами на одном и том же аппаратном обеспечении. Основной вывод Veritest, приведенный в отчете, гласит, что Web-сервер IIS 6.0 на базе Windows Server 2003 позволяет обслуживать большее число одновременно подключенных пользователей, чем Apache 1.3.23 на базе Red Hat Linux Advanced Server 2.1 и Red Hat Linux 8.0 на тестируемом оборудовании, предоставленном компанией Hewlett Packard.

При этом на тестах статических Web-страниц преимущество платформы Microsoft над Unix-платформами составляло от 13% на однопроцессорном сервере HP DL760 до 300% на восьмипроцессорном сервере HP DL760 (рис. 1.10).

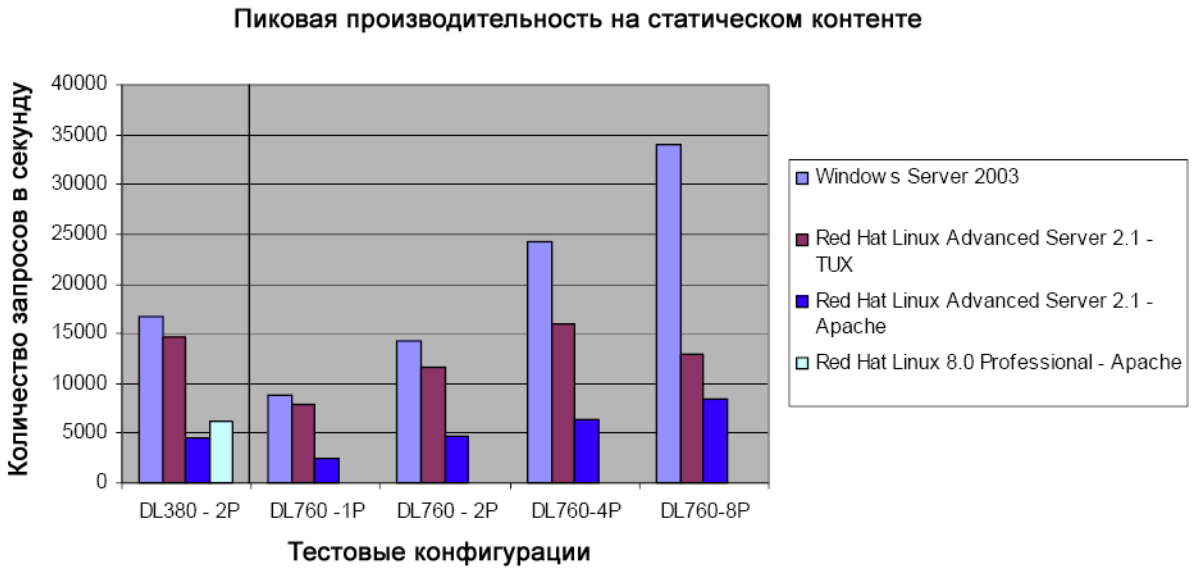

Рис. 1.10. Сравнение производительности на статических Web-страницах [103].

На тестах динамических Web-страниц преимущество платформы Microsoft над Unix-платформами проиллюстрировано на следующей диаграмме (рис. 1.11).

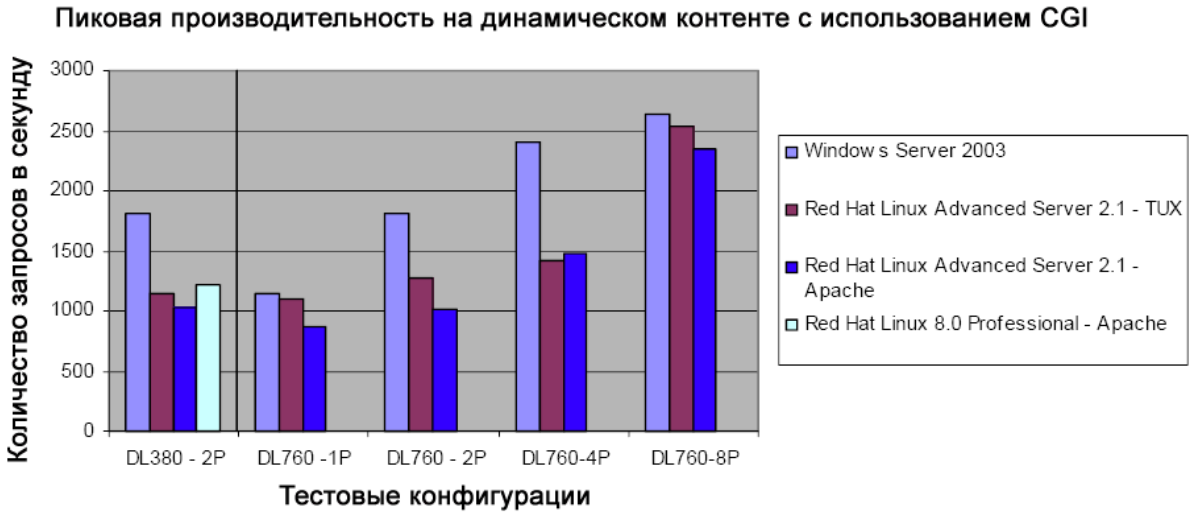

Рис. 1.11. Сравнение производительности на динамических CGI Web-страницах [103].

Следует отметить, что вышеприведенная диаграмма (рис. 1.11) 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. 1.12).

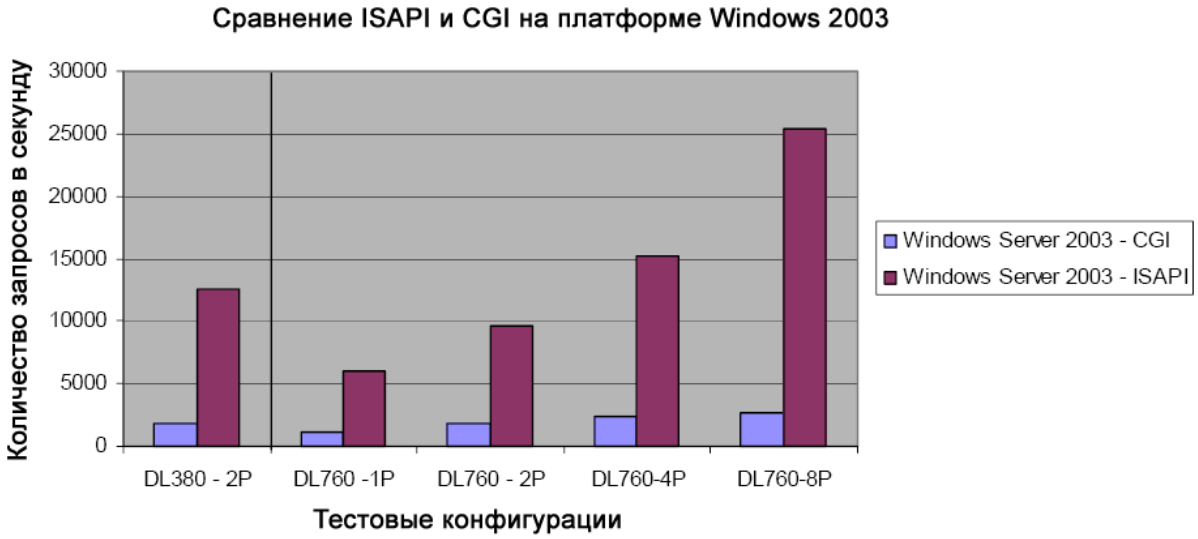

Рис. 1.12. Сравнение производительности ISAPI и CGI на платформе Windows 2003 + IIS 6.0 [103].

Как видно из результатов сравнения проведенного Veritest, платформа Windows 2003 Server + IIS 6.0 является наиболее производительной при построении информационных систем на основе Web-приложений.

1.3.2. 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 марте 2004 года известная аналитическая компания Forrester Research (www.forrester.com) опубликовала доклад, в котором проводится обзор информационной безопасности программного обеспечения, предоставляемого ведущими производителями. В докладе рассматриваются информационные уязвимости, выявленные в период с 1 июня 2002 года по 31 мая 2003 года, и степень успешности их устранения [104].

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. 1.13).

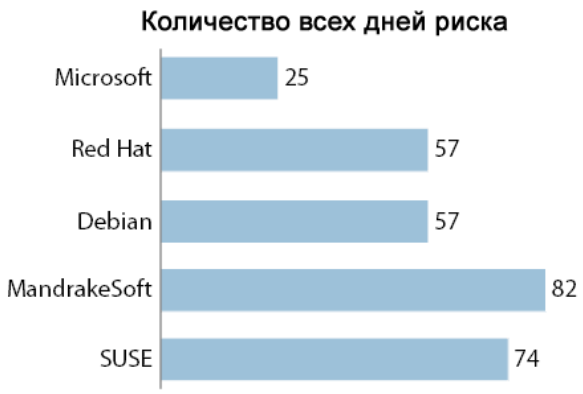

Рис. 1.13. Среднее время (в днях), требующееся на устранение уязвимостей платформ ведущих поставщиков ПО [104].

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 систем и процент их успешного устранения приведен в таблице (табл. 1.2).

Таблица 1.2. Общее число и процент устраненных уязвимостей [104].

Платформа	**Общее кол-во уязвимостей**	**Кол-во критич. уязвимостей**	**% критич. уязвимостей**	**Кол-во испр. уязвимостей**	**% испр. уязвимостей**
Microsoft	128	86	67%	128	100,0%
Red Hat	229	128	56%	228	99,6%
Debian	286	162	57%	275	99,2%
MandrakeSoft	199	120	60%	197	99,0%
SUSE	176	111	63%	172	97,7%

1.3.3. 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 [105].

Вопросы оперативности устранения уязвимостей в коде операционной системы были рассмотрены выше, при рассмотрении информационной безопасности. Что касается времени вынужденного простоя систем на базе Windows и Unix, то этот вопрос проанализирован в исследовании “Microsoft Windows Server 2003 vs. Red Hat

Enterprise Linux AS 3.0: IT Professionals Running a Production Environment”, проведенном компанией Veritest (www.veritest.com) в апреле 2005 года [106]. 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. 1.14).

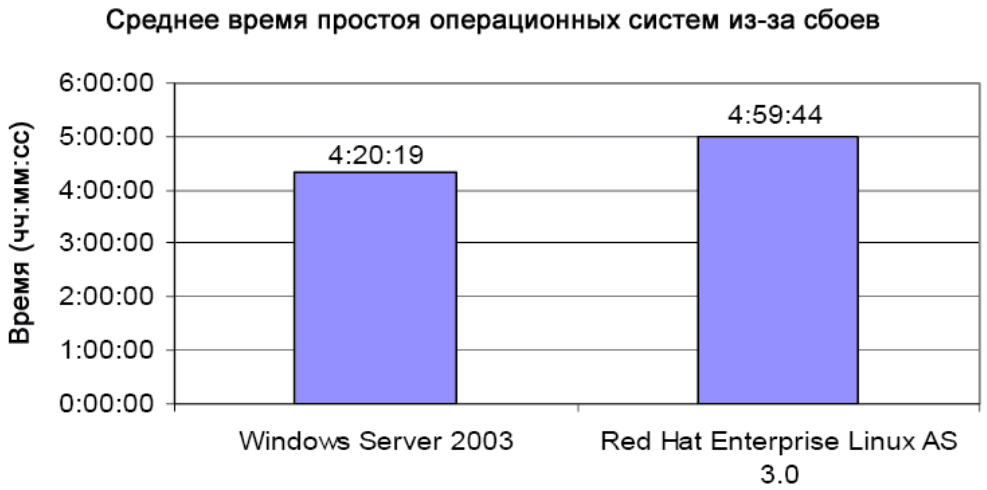

Рис. 1.14. Общее время простоя операционных систем [106].

1.3.4. 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 [107]. 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 [108].

Согласно исследованию “Interoperability: How Technology Managers Rate Microsoft and Its Technologies for Development”, проведенному в апреле 2004 года фирмой Jupiter Research [109], 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.

Примерно в то же время (в июле 2004 года) компания ObjectWatch провела исследование возможности взаимодействия информационных сред с использованием архитектуры, ориентированной на сервисы (SOA) [110]. 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.

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 [111] при исследовании влиятельности основных поставщиков программного обеспечения (рис. 1.15).

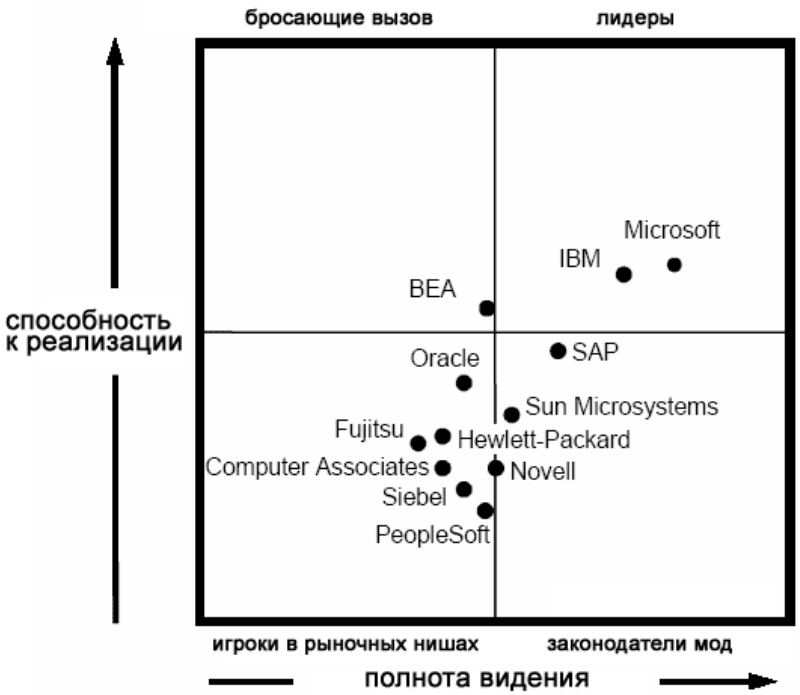

Рис. 1.15. “Волшебный квадрат” влияния основных поставщиков Web-сервисов, полученный Gartner Group в сентябре 2003 года [111].

1.3.5. 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 экономического эффекта при разработке и развертывании приложений на платформах Microsoft и J2EE/Linux, проведенного Forrester Research в сентябре 2003 года [112]. По сути, в этом исследовании сравнивается стоимость реализации ИТ-проектов на базе J2EE/Linux и Microsoft Windows, включая трехгодичную поддержку и сопровождение. Так ССВ для J2EE/Linux составила $2.289.041, в то время как для Microsoft Windows только $1.643.112, что на 28,2% меньше (рис. 1.16).

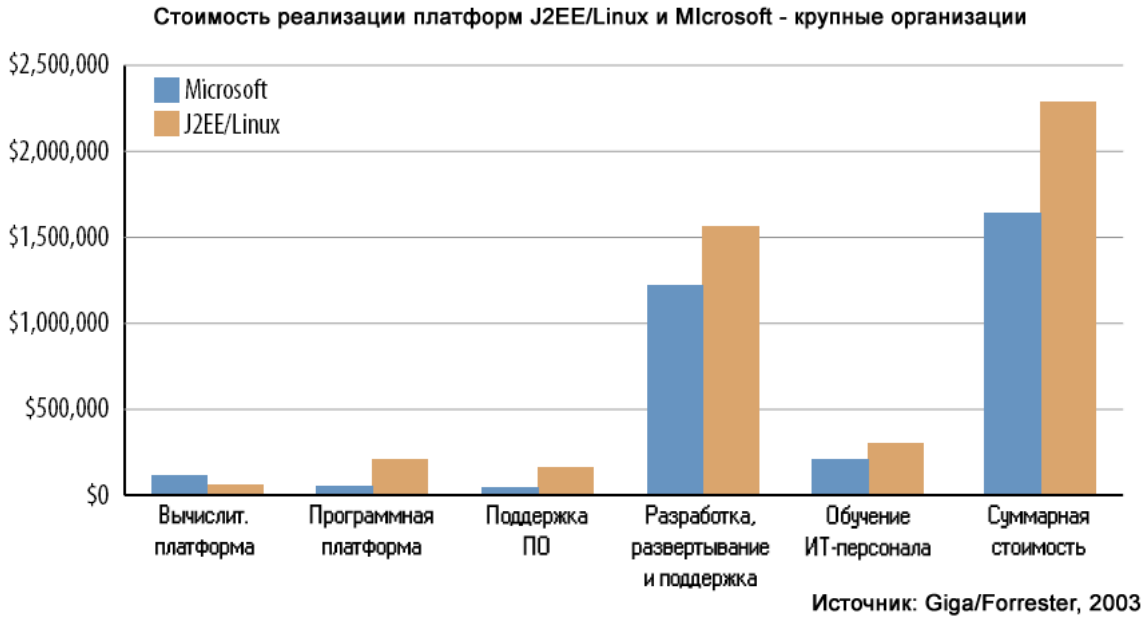

Рис. 1.16. Исследования ССВ решений на базе J2EE/Linux и Microsoft Windows [112].

Если рассматривать только стоимость лицензирования операционных систем их поддержки, то и тут в долгосрочной перспективе продукты Microsoft при полной поддержке 24x7 (24 часа в сутки 7 дней в неделю) оказываются более привлекательными, чем Unix-среды при ограниченной поддержке 9x5 (9 часов поддержки только по будням, рис. 1.17). Данные приведены по результатам исследования, проведенного компанией Bearing Point в мае 2004 года [113].

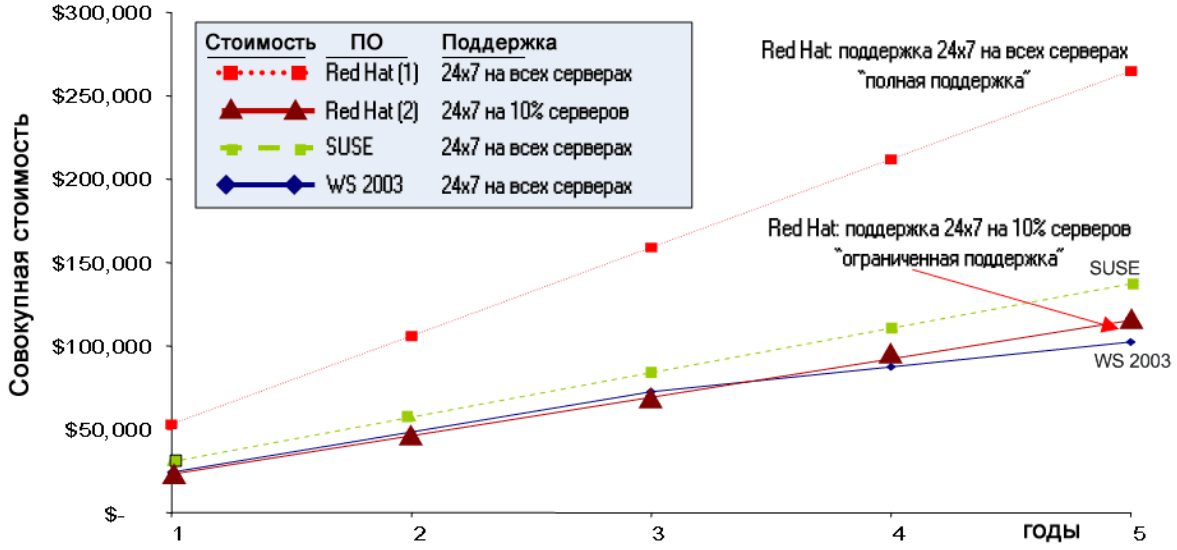

Рис. 1.17. Сравнительная стоимость лицензирования и поддержки Windows и Unix [113].

При сравнении мощных промышленных СУБД (Microsoft SQL Server, IBM DB2 и Oracle 10g) в тестах TPC-C по соотношению цена/производительность платформа на базе Microsoft SQL Server и Windows – явный лидер (табл. 1.3). Данные по состоянию

на 1 ноября 2005 года взяты с Web-сайта Transaction Processing Council [114] – независимой некоммерческой организации, измеряющей производительность СУБД по установленным методикам.

Таблица 1.3. Десятка лучших систем в тестах TPC-C по соотношению цена/производительность [114].

Место	Компания	Система	Произв.	Цена/ Произв.	Доступность системы	СУБД	ОС	Монитор транзакций	Дата тестирования	Кластер
1	DELL	PowerEdge 2800/1/2.8GHz/2+2M	38,622	0.99 USD	11/08/05	Microsoft SQL Server 2005 x64 Std. Ed.	Microsoft Windows 2003 x64 Server Std. Ed.	Microsoft COM+	09/26/05	НЕТ
2	DELL	PowerEdge 2800/1/3.4GHz/2M	28,122	1.40 USD	04/30/05	Microsoft SQL Server 2000 Workgroup Ed.	Microsoft Windows Server 2003 Server	Microsoft COM+	02/24/05	НЕТ
3	DELL	PowerEdge 2850/1/3.4GHz/1M	26,410	1.53 USD	12/10/04	Microsoft SQL Server 2000 Standard Ed.	Microsoft Windows Server 2003 Server	Microsoft COM+	12/10/04	НЕТ
4	hp invent	HP ProLiant ML350T03 X3.06/533/512K SA641	17,810	1.57 USD	10/19/04	Microsoft SQL Server 2000 Standard Ed. SP3	Microsoft Windows Server 2003 Standard Edition	Microsoft COM+	10/19/04	НЕТ
5	hp invent	HP ProLiant ML350-T03-X3.06/533	18,661	1.61 USD	12/15/04	IBM DB2 UDB Express Edition v8.1	SUSE LINUX Enterprise Server 9	Microsoft COM+	07/16/04	НЕТ
6	hp invent	HP Proliant ML350-T03-X2.8/533	18,318	1.68 USD	04/14/04	IBM DB2 UDB Express Edition v8.1	Microsoft Windows Server 2003 Standard Edition	Microsoft COM+	04/14/04	НЕТ
7	hp invent	HP ProLiant ML370-G4-1M 3.6Ghz 2P	68,010	1.80 USD	11/01/04	Microsoft SQL Server 2000 Enterprise Ed. SP3	Microsoft Windows Server 2003 Enterprise Edition	Microsoft COM+	11/01/04	НЕТ
8	hp invent	HP Integrity rx2600	51,506	1.81 USD	09/29/04	Oracle Database 10g Standard Edition	Red Hat Enterprise Linux AS 3	Microsoft COM+	09/29/04	НЕТ
9	hp invent	HP ProLiant ML370-G3-2M/3.20GHz-1P	35,030	1.88 USD	03/17/04	Microsoft SQL Server 2000 Enterprise Edition	Microsoft Windows Server 2003 Enterprise Edition	Microsoft COM+	03/02/04	НЕТ
10	hp invent	HP Proliant ML350T03 X3.06/533	17,192	1.96 USD	12/17/03	Microsoft SQL Server 2000 Standard Ed. SP3	Microsoft Windows Server 2003 Standard Edition	Microsoft COM+	12/17/03	НЕТ

Если рассматривать тесты TPC-W, которые оценивают результаты в транзакционных системах, основанных на Web по соотношению цена/производительность, то эти данные красноречиво показывают первенство Microsoft (табл. 1.4).

Таблица 1.4. Пятерка лучших систем в тестах TPC-W по соотношению цена/производительность [114].

Место	Компания	Система	Произв.	Цена / Произв.	Доступность системы	СУБД	ОС	HTTP-сервер	Дата тестирования
1	DELL	PowerEdge 6650	9,708	34.60 USD	05/31/02	Microsoft SQL Server 2000 Enterprise Edition	Microsoft Windows 2000 Advanced Server	Microsoft Internet Information Server 5.0	05/31/02
2	IBM	IBM eServer xSeries 370 with xSeries 330	6,045	76.67 USD	05/31/01	Microsoft SQL Server 2000	Microsoft Windows 2000 Advanced Server	Microsoft Internet Information Server 5.0	05/31/01
3	UNISYS	e-@ction Enterprise Server ES 7000 (16P)	10,439	106.73 USD	07/10/01	Microsoft SQL Server 2000 Enterprise Edition	Microsoft Windows 2000 Datacenter Server	Microsoft Internet Information Server 5.0	07/10/01
4	IBM	IBM eServer xSeries 430 with IBM Netfinity 4500R	7,554	136.80 USD	06/08/01	IBM DB2 UDB 7.2	IBM ptx 4.6.1	Microsoft Internet Information Server 5.0	05/01/01
5	IBM	NUMA-Q E410	6,272	195.59 USD	02/02/01	IBM DB2 UDB 7.1	DYNIX/ptx 4.5.1	Microsoft Internet Information Server 5.0	02/02/01

1.3.6. 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 первой главе получены следующие результаты:

- на основе анализа различных способов интеграции гетерогенных информационных систем для интеграции информационных источников выбран подход Local-As-View (Enterprise Information Integration);
- рассмотрены конфликты гетерогенности, которые необходимо разрешить при разработке интегрированной информационной системы;
- рассмотрены наиболее значимые базы данных по свойствам неорганических веществ, созданные в мире;
- проанализированы современные программные платформы для построения интегрированной информационной системы, выработаны критерии отбора для решения поставленных задач и на их основе сделан выбор в пользу платформы Microsoft.

ГЛАВА 2. МЕТОДИКА ПОСТРОЕНИЯ ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ ПО СВОЙСТВАМ ВЕЩЕСТВ ДЛЯ ЭЛЕКТРОНИКИ

Для того чтобы выработать методику построения интегрированной информационной системы (ИС) по свойствам веществ для электроники, 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, равно как и без их семантического понимания, обречена на провал.

2.1. Информационные системы по свойствам неорганических веществ ИМЕТ РАН

Как было отмечено в разделе 1.2, в настоящее время в мире существует множество ИС по свойствам веществ [4-18, 45-101]. При этом в каждой исследовательской организации создавались свои собственные информационные ресурсы, в которых накапливалась информация, относящаяся к её 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.

Таким образом, ИС ИМЕТ РАН являются репрезентативными для проведения анализа их построения и разработки некоторой общей схемы интегрируемых ИС, в которую бы вписывались все указанные ИС. Проанализируем некоторые из ИС ИМЕТ РАН:

- ИС по свойствам неорганических соединений “**Фазы**”;
- ИС по фазовым диаграммам систем с полупроводниковыми фазами “**Диаграмма**”;
- ИС по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами “**Кристалл**”;
- ИС по ширине запрещенной зоны неорганических соединений “**BandGap**”;
- ИС по свойствам полупроводникового кремния “**Кремний**”.

Следует отметить, что ИС “Кристалл” и “BandGap” были разработаны нами при проведении исследований в рамках диссертационной работы.

2.1.1. 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 соединения 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).

2.1.2. ИС по фазовым диаграммам систем с полупроводниковыми фазами “Диаграмма”

Основной информацией ИС “Диаграмма” [46-48, 52, 53] 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 [46]. В настоящее время БД содержит информацию о нескольких десятках двойных и тройных систем, извлеченную из около 2 тыс. публикаций. Содержимое БД постоянно пополняется новыми данными [115, 116].

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 jpeg-формате [117]. 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).

2.1.3. ИС по веществам с особыми акустооптическими, электрооптическими и нелинейнооптическими свойствами “Кристалл”

В рамках диссертационных исследований нами разработана ИС “Кристалл”, построенная на основе БД Microsoft SQL Server 2000, содержащей экспериментальные данные о свойствах акустооптических, электрооптических и нелинейнооптических

веществ, погрешностях, методах измерений, условиях получения и т.д. [50, 51]. 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 о свыше ста 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.

При разработке ИС “Кристалл” была реализована возможность многокритериального поиска данных. Например, пользователь может вывести информацию обо всех электрооптических веществах, имеющих заданную сингонию, плавящихся без разложения ниже заданной температуры и имеющих при заданной длине волны значение определенного линейного электрооптического коэффициента в заданном интервале. Для этого ему необходимо только отметить выбранные свойства в перечне свойств веществ, представленных в ИС, и последовательно ввести критерии поиска. Стоит отметить, что в настоящее время разработаны русскоязычная и англоязычная версии ИС “Кристалл”. Информация в БД ИС обновляется с использованием специализированного ПО [115, 118].

ИС была разработана в виде Web-приложения, построенного на технологии Active Server Pages (ASP). Для доступа к БД, работающей под управлением Microsoft

SQL Server 2000, используется интерфейс ActiveX Data Objects (ADO). Web-приложение ИС развернуто на Web-сервере Microsoft IIS 6.0 под управлением Microsoft Windows 2003 Server в ИМЕТ РАН. Для зарегистрированных пользователей ИС “Кристалл” доступна из глобальной сети Интернет (http://crystal.imet-db.ru).

2.1.4. ИС по ширине запрещенной зоны неорганических соединений “BandGap”

В рамках диссертационных исследований нами разработана ИС “BandGap”, которая содержит информацию по ширине запрещенной зоны (Eg) основных неорганических соединений. 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 HF, энтропии ΔSm и температуры плавления Tm и других) [119, 120].

В ИС “BandGap” специально отмечены приводимые оригинальными авторами 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.

Основу ИС составляет реляционная БД под управлением Microsoft SQL Server 2000, содержащая описания неорганических соединений, численные данные по ширине запрещенных зон неорганических соединений и ссылки на библиографические источники и дополнительные графические материалы. ИС разработана нами в виде Web-приложения, реализованного с использованием технологии Active Server Pages (ASP). Для доступа к данным, хранимым на Microsoft SQL Server 2000, используется интерфейс ActiveX Data Objects (ADO). Web-приложение ИС развернуто на Web-сервере Microsoft IIS 6.0 под управлением Microsoft Windows 2003 Server в ИМЕТ РАН. Для зарегистрированных пользователей ИС “BandGap” доступна из глобальной сети Интернет (http://bg.imet-db.ru).

2.1.5. 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 веществом, 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, 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).

2.1.6. Разработка программного комплекса для удаленного администрирования произвольных 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 вы должны написать по сравнению с функциональностью, которую вы получаете от него.

В табл. 2.1 содержатся результаты сравнительного анализа сильных и слабых сторон различных интерфейсов доступа к базам данных. Приняты следующие условные обозначения: знак “+” обозначает сильную сторону, “++” обозначает особенную мощь,

“–” обозначает слабую сторону, и, наконец, отсутствие знаков означает отсутствие значимых достоинств и недостатков.

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 возможных интерфейсов для реализации программного комплекса.

Таблица 2.1. 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”	–		+		–	+	+

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 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 является, в нашем случае, 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. 2.1). Комплекс разработан на языке C++ с использованием RAD-среды Borland C++Builder 6 Enterprise Edition [116]. 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.

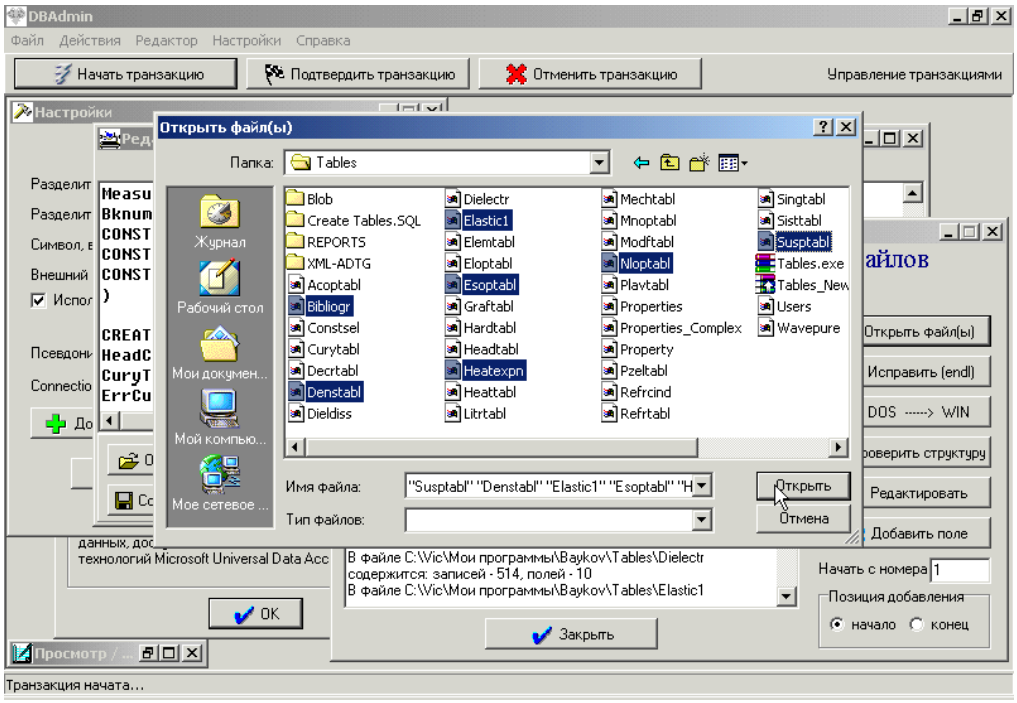

Рис. 2.1. Интерфейс программного комплекса DBAdmin.

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.

2.1.7. 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).

2.2. 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 являлась задача отображения рисунков фазовых диаграмм (ФД) для конечного пользователя ИС. Отметим, что возможны различные методы представления рисунков ФД [121]:

- использование термодинамических моделей элементов ФД с возможностью последующей визуализации рассчитанных кривых, что влечет за собой написание расчетно-визуализирующей подсистемы [122-124];
- использование различных аппроксимирующих математических моделей элементов ФД (обычно полиномов) также с последующей визуализацией рассчитанных кривых (написание расчетно-визуализирующей подсистемы) [122, 125-127];
- 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) [127-129];
- 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) [123, 126].

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 [121, 128], 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) [117], 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) [130]. 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 произведена 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. 2.2 приведен пример выдачи графической информации из ИС “Диаграмма”.

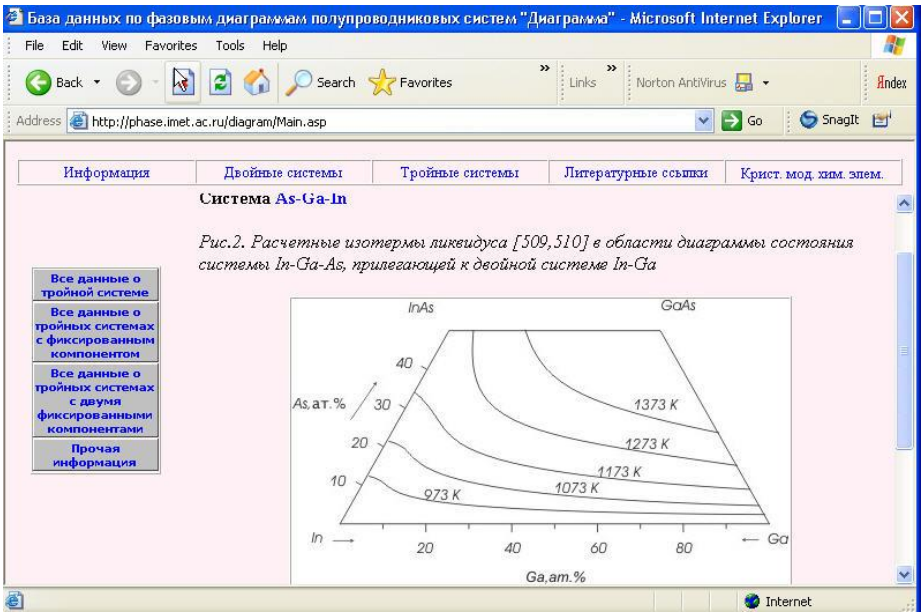

Рис. 2.2. Пример выдачи графической информации из БД “Диаграмма”.

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 [94], 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, которое 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.

2.3. Архитектура современных информационных систем по свойствам веществ

Рассмотренные ИС ИМЕТ РАН спроектированы с учетом всех современных требований, предъявляемых к построению ИС. Так, все эти системы располагают Web-интерфейсами и доступны для зарегистрированных пользователей через глобальную сеть Интернет. Если рассматривать архитектурно-технологические аспекты построения этих систем, то можно обнаружить много схожего.

При разработке программного обеспечения (ПО) всех 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. Трёхзвенная архитектура – вариант архитектуры клиент-сервер, в которой пользовательский 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). Стрелками зеленого цвета на рисунке отображены потоки информации между звеньями.

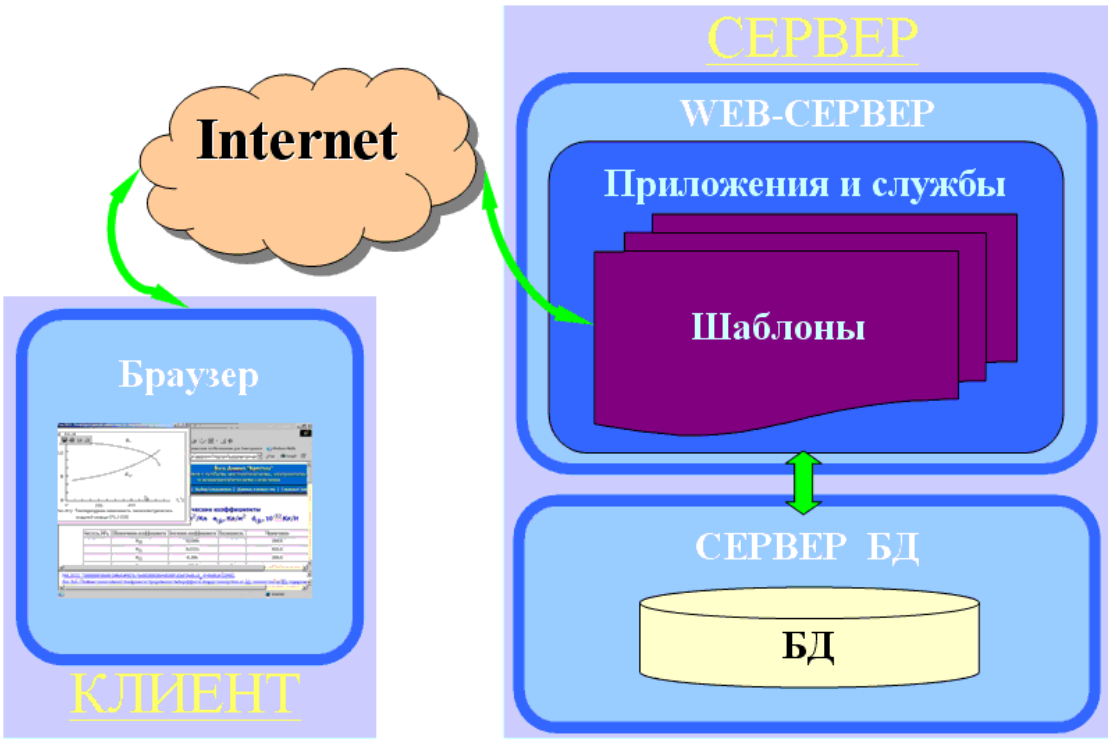

Рис. 2.3. Трехзвенная архитектура ИС с доступом пользователей через Интернет.

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 [131].

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 всех ИС ИМЕТ РАН ставилась задача обеспечения максимально простого программного обеспечения на клиентской части, то принималось решение об использовании Интернет-браузера. Интернет-браузер является стандартным компонентом ПО в современных персональных компьютерах и, как правило, входит в состав наиболее распространенных операционных систем (ОС). Например, во все ОС корпорации Microsoft по умолчанию встраивается браузер Microsoft Internet Explorer. Таким образом, от пользователя ИС не требуется устанавливать дополнительное ПО, обеспечивающее взаимодействие с рассматриваемыми ИС. Пользователю достаточно иметь любой браузер 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, как правило, 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 [132], 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 активно используется в ИМЕТ РАН, для повышения производительности и устойчивости функционирующих ИС.

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 2000 (под управлением ОС Windows Server 2003) и Oracle 8*i* (под управлением ОС Sun Solaris 2.5).

В качестве звена обработки информации в ИС ИМЕТ РАН выступает Web-сервер. 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 виртуальных машин, так и обеспечение работы одного Web-приложения распределенно сразу на нескольких Web-серверах (так называемая Web-

factory). Это значительно повышает масштабируемость и отказоустойчивость современных Web-приложений. В качестве Web-серверов в ИМЕТ РАН используются сервера на базе технологий Microsoft (это Microsoft Internet Information Services (MS IIS)). 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.

2.4. Комплексный подход к интеграции информационных систем

Отличительной особенностью информационных систем по свойствам неорганических веществ, 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 их состав, 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 в виде дерева (рис. 2.4):

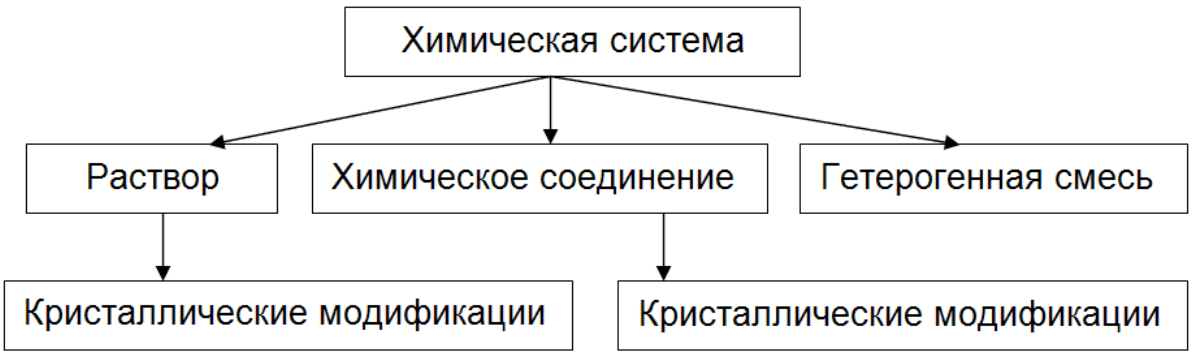

Рис. 2.4. Иерархия химических понятий.

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 [133]:

Химическая система – система, образованная химическими элементами. Она может быть описана как множество обозначений химических элементов, образующих химическую систему. 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 от свойств образующих его элементов химическим или кристаллохимическим строением. Соединение образовано из атомов нескольких химических элементов, связанных химической связью. На фазовой диаграмме область гомогенности соединения отделена (при всех температурах и давлениях) от области компонентов или твердых растворов на их основе. Элементы в соединении не могут быть разделены простым механическим способом, а лишь химической обработкой, нагреванием или электрическим током.

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.

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;
- уровень химических веществ;
- уровень 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 [94], 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 основных подхода к интеграции современных информационных ресурсов по свойствам веществ [134]. Первый подход заключается в интеграции информационных ресурсов на уровне виртуального объединения 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 [135]. 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).

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 систем [136, 137]. Таким образом, второй подход 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 соответствующих информационных систем. Этот подход, 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.

Особо стоит отметить, что при использовании обоих подходов осуществляется "виртуальное" интегрирование ИС по свойствам веществ и создание неоднородной распределенной информационной системы. Оба рассмотренных подхода подразумевают независимость развития отдельных подсистем, при этом пользователь получает доступ ко всему массиву данных о конкретном веществе или соединении, хранящемуся в информационных системах виртуальной объединенной системы, что и решает основную задачу интеграции. При этом в первом случае (при интеграции структурированной информации на уровне источников данных) пользователь получает

доступ к интегрированной по соответствующим схемам информации, а во втором случае (при интеграции пользовательских интерфейсов информационных систем) пользователь получает доступ к богатым возможностям информационно-расчетных подсистем конкретной интегрированной ИС.

Таким образом, при нынешних условиях развития современных ИС по свойствам веществ, необходим комплексный подход к интеграции. Это означает необходимость использования интеграции информационных систем как на уровне информации (в разделе 1.1 был выбран подход Local-As-View EII), содержащейся в источниках данных ИС, так и на уровне пользовательских интерфейсов ИС (подход EAI). При этом следует особо отметить, что обеспечивается независимость развития отдельных ИС, так как созданные ИС продолжают свое развитие, пополняясь информацией и расширяясь новыми расчетными подсистемами, обеспечивающими дополнительные функциональные возможности ИС. Предложенный комплексный подход позволяет успешно сочетать в себе все достоинства разработанных и продолжающих эволюционировать информационных систем по свойствам веществ. В рамках разработанного подхода предоставляется как доступ к текущим пользовательским интерфейсам ИС 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.

2.5. 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) [132] и интерфейс CGI (Common Gateway Interface) [138]. 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 “метабаза” – специальная база данных, содержащая справочные сведения о содержимом интегрируемых ИС, а именно – о химических системах и их свойствах (рис. 2.5). Этих сведений достаточно для того, чтобы построить поиск релевантной информации по химическим системам и их свойствам.

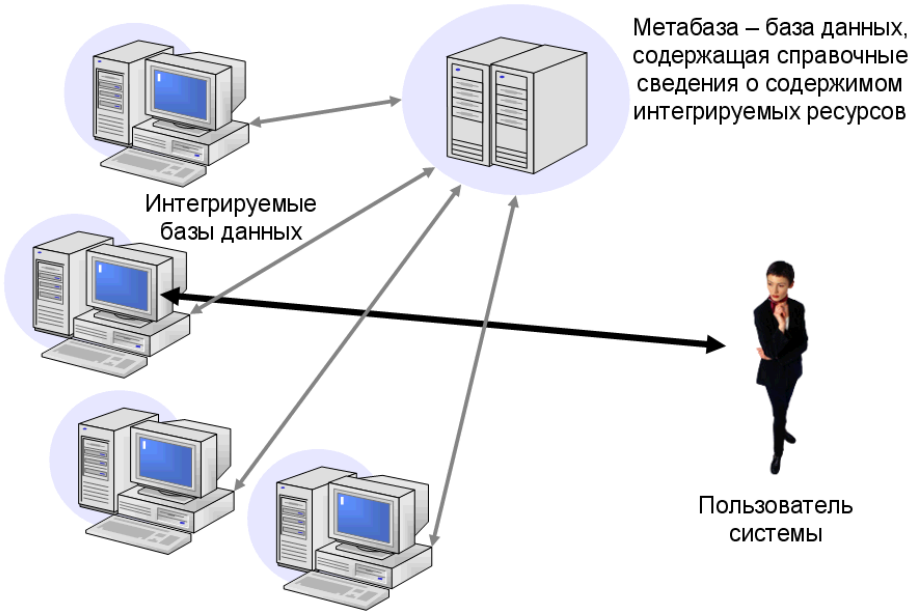

Рис. 2.5. Метабаза как координирующий центр интегрированной ИС.

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:

1. Для любых множеств $s_1 \in S, s_2 \in S$, состоящих из химических элементов e_{ij}, $s_1 = \{e_{11}, e_{12}, .., e_{1n}\}, s_2 = \{e_{21}, e_{22}, .., e_{2m}\}$ верно, что если $s_1 \subseteq s_2$ (то есть, все химические элементы из системы s_1 содержатся в системе s_2), то $(s_1, s_2) \in R$.
2. 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}, $s_1 = \{e_{11}, e_{12}, .., e_{1n}\}, s_2 = \{e_{21}, e_{22}, .., e_{2m}\}$ 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 целесообразным 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:

$R \subset (d_1, s_1, p_1) \times (d_2, s_2, p_2)$, 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.

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. 2.6). 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.

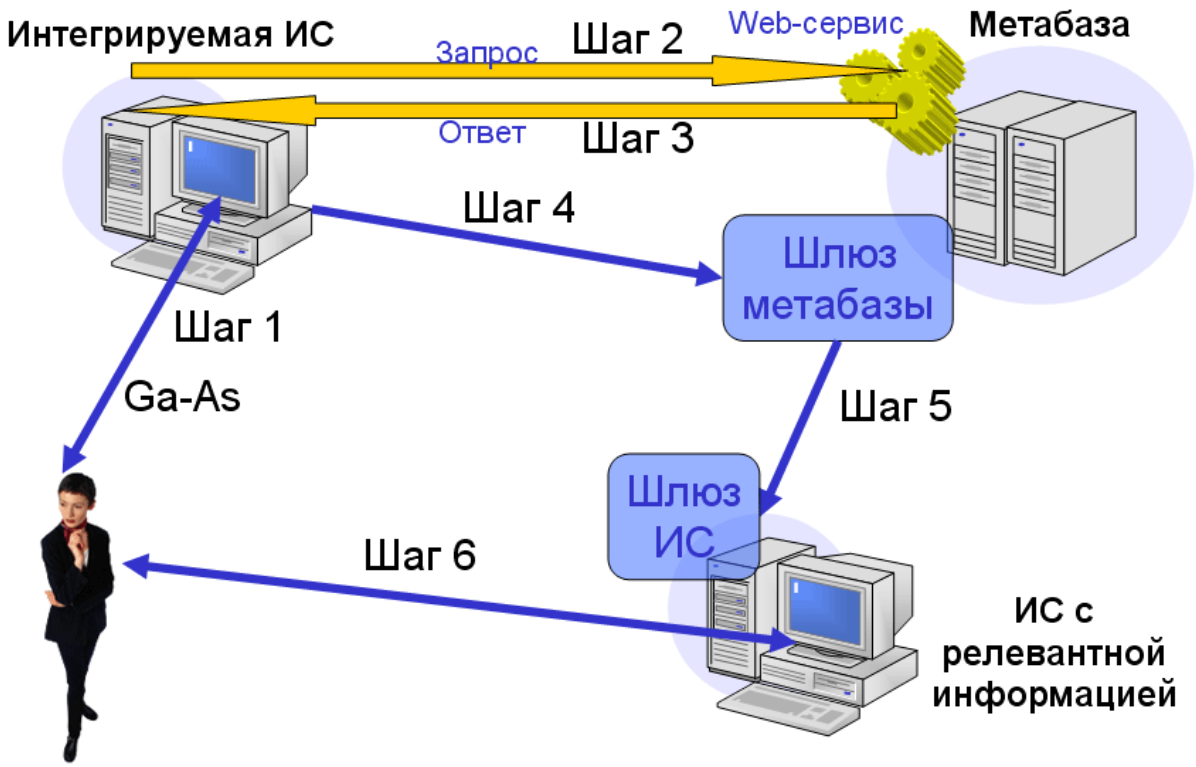

Рис. 2.6. Интеграция Web-приложений ИС с поиском релевантной информации в метабазе.

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. 2.6. 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. 2.6). 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. 2.6). 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. 2.6). 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 [139].

В нашем случае, целесообразно выводить релевантную информацию в виде гиперссылок, щелкая 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, щелкая по ссылке на релевантную информацию, должен сначала попадать на шлюз метабазы (шаг 4 на рис. 2.6).

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. 2.6).

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.

2.6. 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. раздел 2.5), где применяется EAI-интеграция, при объединении гетерогенных источников данных речь идет уже о EII-интеграции (выбранной в разделе 1.1). 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.

Для проведения интеграции на уровне источников данных в разделе 1.1 был выбран подход Local-As-View в рамках EII (EII LAV), который позволяет строить масштабируемые интегрированные ИС. Основой при построении интегрированных ИС в рамках данного подхода является разработка общей глобальной схемы предметной области. Следует отметить, что в отличие от подхода Global-As-View, при подходе Local-As-View глобальная схема должна быть особенно тщательно продуманной, поскольку именно в ее терминах будут представлены все интегрируемые источники информации. 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.

Как было отмечено в разделе 1.1, при проведении интеграции источников информации 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 по свойствам неорганических веществ.

2.6.1. 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 (например, Microsoft Windows, Unix и т.д.), СУБД (например, Microsoft SQL Server, Oracle и т.д.) и другого ПО (например, Web-сервера Microsoft IIS, Apache 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 [140].

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 [141].

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 [142]).

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.

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 [140].

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.

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).

2.6.2. 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 2.4, при описании химических объектов можно использовать три уровня: система, вещество и кристаллическая модификация (далее, просто, модификация). Указанная иерархия химических объектов, которая будет рассматриваться в контексте интегрированной ИС, представлена на рис. 2.7.

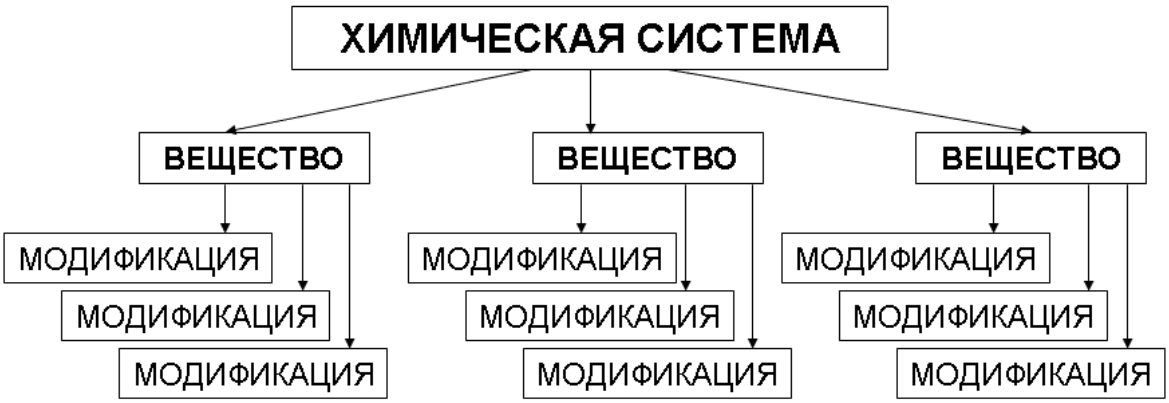

Рис. 2.7. Иерархия химических объектов, рассматриваемая в контексте интегрированной ИС по свойствам неорганических веществ.

Таким образом, в общую схему предметной области закладывается три типа объектов, соответствующих химическим сущностям: система, вещество 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).

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: $s = \{e_1, e_2, .., e_n\}$. 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 $< s, f, \mathrm{mod} >$, 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 сущностей 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 сущностей.

Как было отмечено, в интегрируемых ИС содержится информация по свойствам химических сущностей. 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 [143]. 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 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.

2.6.3. 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 идентификаторов. 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. 2.8).

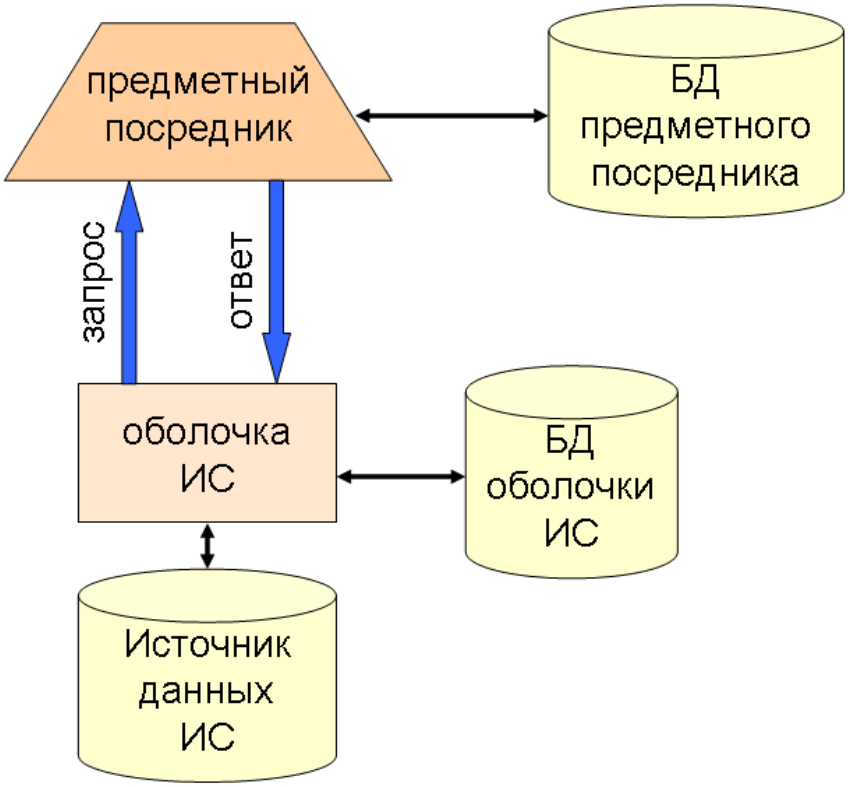

Рис. 2.8. Схема разрешения семантических конфликтов в интегрированной ИС.

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 сущности 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: $s = \{e_1, e_2, .., e_n\}$. После формирования XML-документа, оболочка ИС посылает запрос Web-сервису предметного посредника. Предметный посредник анализирует каждую химическую систему $s = \{e_1, e_2, .., e_n\}$ 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 концентрацию 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 нестандартизировано в разных ИС. В результате возникают трудноразрешимые семантические конфликты при обозначении одних и тех же модификаций различными строковыми литералами. 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 “тр”. 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 подход к интеграции, основанный на принципах EAI. Параллельное применение подходов EAI и EII позволит построить не имеющую аналогов интегрированную ИС по свойствам неорганических веществ. Данная ИС сможет широко использоваться не только специалистами в области химии, но и системами поддержки принятия решений (СППР) для анализа данных в ИС, нахождения закономерностей и построения гипотез о существовании веществ с заданными характеристиками.

Краткие выводы

Во второй главе получены следующие результаты:

- на основе проведенного анализа архитектуры современных информационных систем принято решение использовать оригинальный комплексный подход к интеграции информационных систем. Интеграцию решено проводить на уровне Web-интерфейсов ИС с использованием интеграционного подхода Enterprise Application Integration и на уровне источников информации с использованием подхода Local-As-View (Enterprise Information Integration);
- формализовано понятие релевантной информации, содержащейся в интегрируемых информационных системах рассматриваемой предметной области;
- разработана методика интеграции распределенных Web-приложений информационных систем с учетом требований информационной безопасности и поисковым механизмом для обнаружения релевантной информации, содержащейся в рамках интегрируемых информационных систем;
- на основе анализа информационных систем по свойствам веществ формализована глобальная схема предметной области для интеграции информационных источников, основанная на иерархии химических понятий, которую предлагается использовать в рамках интегрированной информационной системы;
- разработана методика интеграции гетерогенных информационных источников на основе предложенных механизмов разрешения конфликтов гетерогенности и схемы предметной области.

ГЛАВА 3. РАЗРАБОТКА ПРОГРАММНОГО КОМПЛЕКСА ИНТЕГРИРОВАННОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ

В главе 2 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 [144].

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.

3.1. Реализация интеграции гетерогенных Web-приложений информационных систем

Принципы построения единой информационной системы (ИС), объединяющей распределенные гетерогенные Web-приложения интегрируемых ИС по свойствам неорганических веществ, были рассмотрены в разделе 2.5. В данной части работы рассмотрены важные аспекты реализации интегрированной ИС.

В разделе 1.3. рассматривалась возможность построения ИС на базе доминирующих в настоящий момент технологических Web-платформ Windows/IIS и Unix/Apache. В результате оценки платформ согласно выбранным критериям было принято решение использовать технологические платформы компании Microsoft (Windows/IIS).

3.1.1. 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. Эти данные необходимы для обеспечения поиска релевантной информации согласно методике, описанной в разделе 2.5. Для хранения данных 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 2000.

Структура метабазы во многом определяет функциональные возможности интегрированной ИС, поэтому в настоящем разделе повышенное внимание будет уделено рассмотрению структурных деталей БД, лежащей в основе ИС по объединению Web-приложений интегрируемых ИС [144]. Для решения задачи объединения Web-приложений ИС по свойствам неорганических веществ предлагается следующая структура метабазы (рис. 3.1).

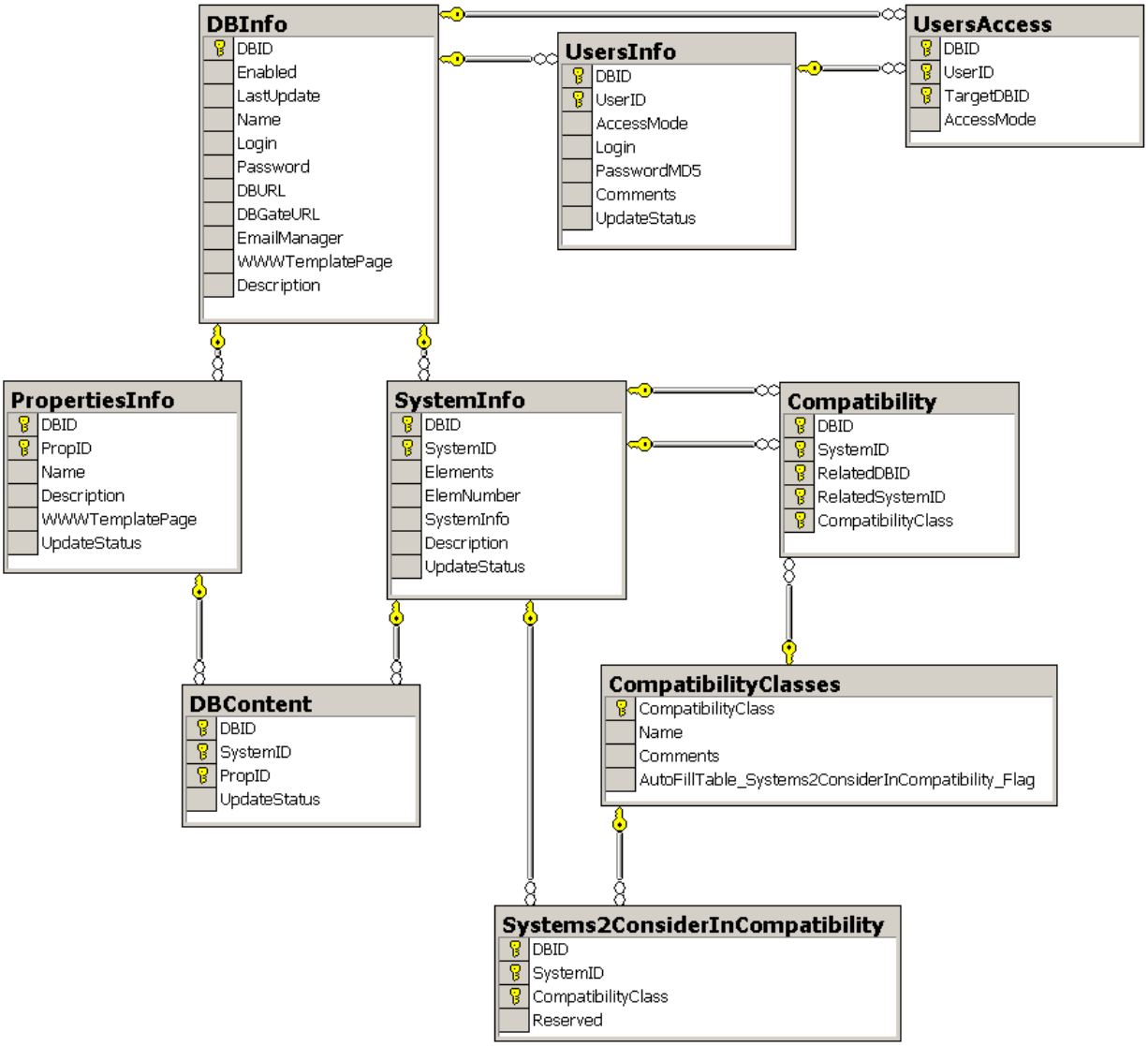

Рис. 3.1. Структура метабазы информационной системы, интегрирующей Web-приложения информационных систем по свойствам неорганических веществ.

Рассмотрим кратко назначение таблиц метабазы для интеграции Web-ресурсов по свойствам веществ, указывая, какую функциональность, опирающуюся на структуру этих таблиц, будет реализовывать интегрированная ИС.

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 2.5. 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 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 2.5. 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 2.5. 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 разделе 2.5, 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 (как было показано в разделе 2.5), или осуществлять их построение вручную. Первичным ключом таблицы является поле CompatibilityClass (тип int), являющееся идентификатором класса релевантности. Поле

AutoFillTable_Systems2ConsiderInCompatibility_Flag (тип int) содержит признак необходимости добавления новых записей химических систем в таблицу Systems2ConsiderInCompatibility 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 CompatibilityClass, 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_Systems2ConsiderInCompatibility_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 Versions

Эта таблица не отображена на рис. 3.1, 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 4). 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;
- **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;
- **Versions** – служебная таблица для организации системы версий ПО.

3.1.2. 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 2.5. 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. 3.2). Разработка велась в среде Microsoft Visual Studio 2003. Рассмотрим кратко механизм обновления метаданных (полностью данный механизм описан в [144]). 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 [142, 145]. 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 [146]. Так в метабазе появляется актуальная информация о содержимом интегрируемых информационных ресурсов.

Рис. 3.2. 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) [147]. 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 изменений. 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 2.5. 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.

3.1.3. 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. 3.3).

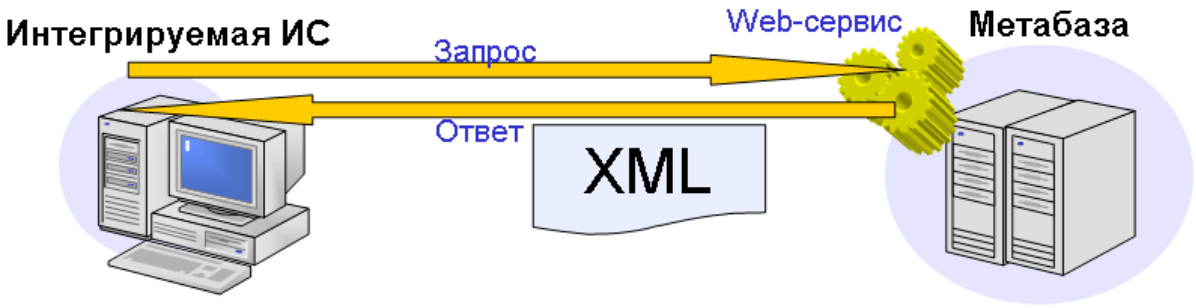

Рис. 3.3. Запрос релевантной информации у Web-сервиса метабазы.

Взаимодействие с Web-сервисом осуществляется по протоколу SOAP согласно WSDL-описанию сервиса [148]. 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 [144].

Основной функцией, поддерживаемой Web-сервисом поиска релевантной информации, является предоставление списка релевантной информации. Данная функциональность поддерживается методом GetIntegratedInfo4UserByCompatibilityClass, 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 [142, 149]. 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. 3.4. Формат этого XML-документа будет подробнее разобран ниже при описании механизма безопасного перехода пользователя из одного Web-приложения в другое.

```
  <?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. 3.4. Пример XML-документа, возвращаемого Web-сервисом поиска релевантной информации (снимок экрана из Microsoft IE).

3.1.4. 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. 3.5. Рассмотрим пошагово процесс перехода пользователя.

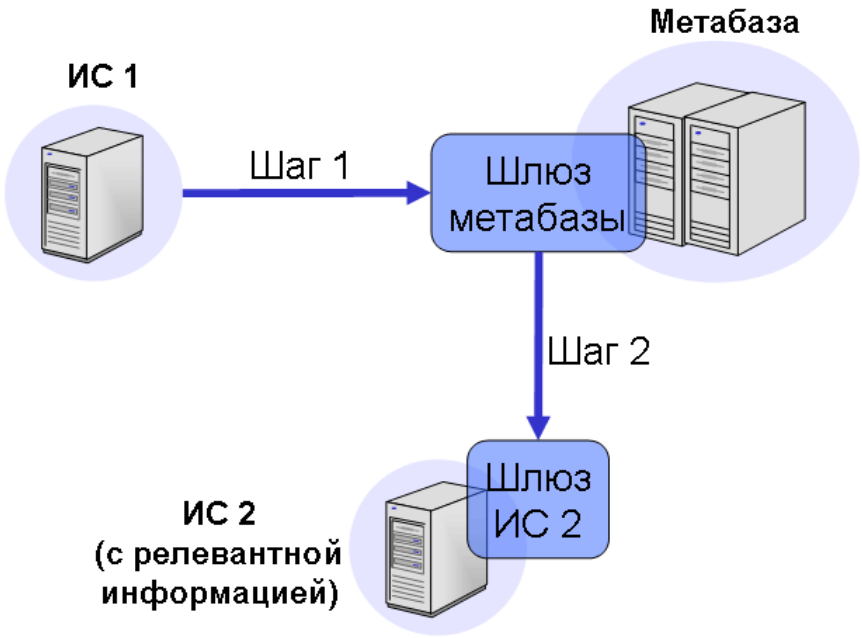

Рис. 3.5. Процесс перехода пользователя между Web-приложениями информационных систем через шлюз метабазы.

Сначала пользователь запрашивает список релевантной информации у Web-сервиса поиска релевантной информации и получает в ответ XML-документ (см. рис. 3.4), 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). 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 [144]:

- **idd** – параметр, содержащий идентификатор информационной системы, пользователь которой осуществляет переход (исходная ИС или “ИС 1” на рис. 3.5);
- **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. 3.5);
- **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;
- **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 на просмотр релевантной информации по указанным параметрам. 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 (шаг 1 на рис. 3.5), щелкнув по соответствующей ссылке в Web-интерфейсе соответствующей ИС. На рис. 3.6 показано как может быть отображена релевантная информация в окне браузера пользователя на примере Web-приложения ИС “BandGap”.

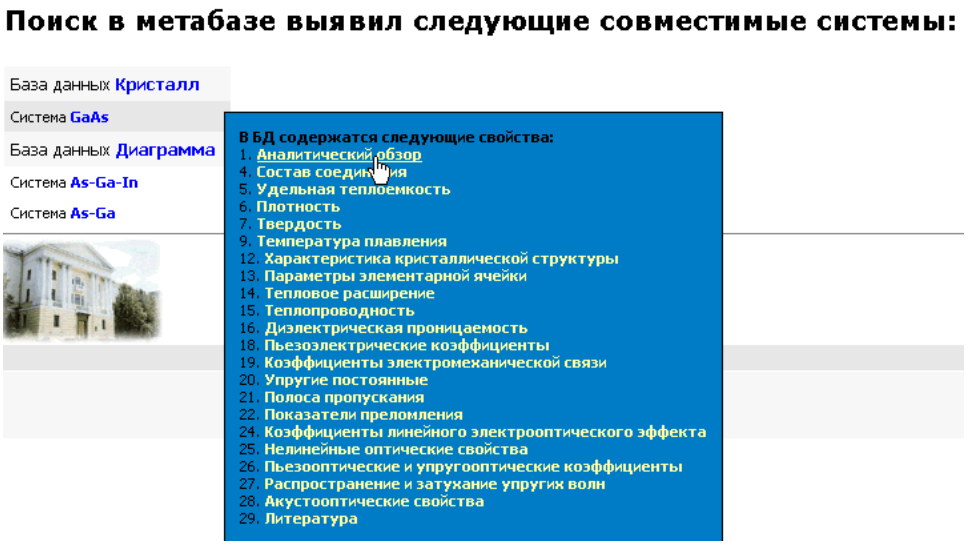

Рис. 3.6. Пример отображения релевантной информации в ИС “BandGap”.

На шлюзе безопасности метабазы происходит анализ параметров запроса [144] 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. 3.5). 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 [144].

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.

В настоящее время интеграция на уровне Web-приложений проведена для ИС “BandGap”, “Кристалл”, “Фазы” и “Elements”. Таким образом, пользователи любой из этих ИС могут просматривать релевантную информацию, содержащуюся в смежных ИС. Планируется функциональное расширение интегрированной ИС за счет подключения новых ИС по свойствам неорганических веществ и введения новых классов релевантности.

3.2. Реализация интеграции гетерогенных источников данных информационных систем

Принципы построения информационной системы, объединяющей гетерогенные источники данных интегрируемых информационных систем (ИС) по свойствам неорганических веществ, были рассмотрены в разделе 2.6. В настоящем разделе 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 разделе 1.3 настоящей работы, было принято решение использовать технологические платформы компании Microsoft (Windows/IIS), что позволит минимизировать затраты на разработку интегрированной ИС.

3.2.1. 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 2.6. 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 2005 (кодовое название Yukon), обладающей богатыми возможностями (.Net CLR, XPath, XQuery) по работе с XML-документами внутри БД.

В связи с тем, что интегрируемые ИС 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.

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. 3.7).

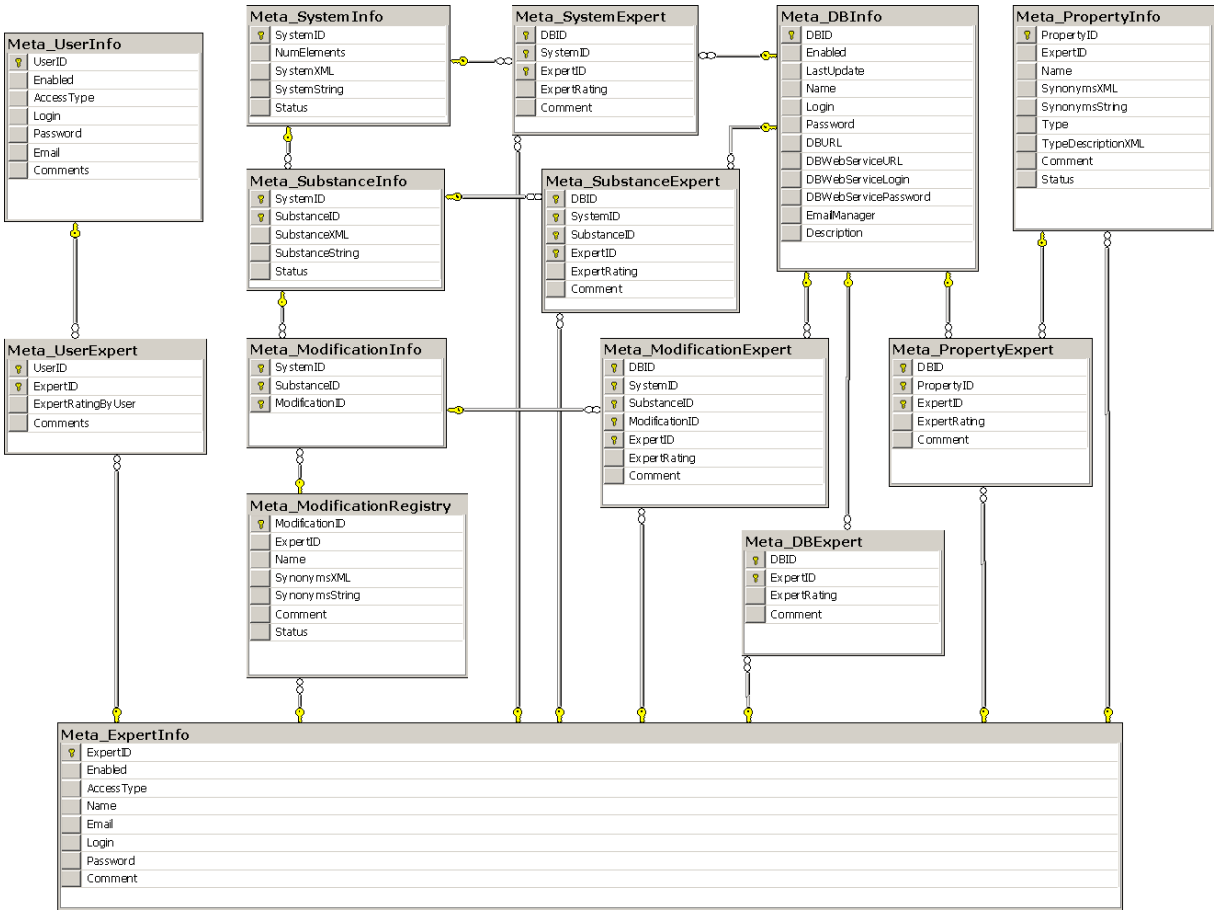

Рис. 3.7. Структура метабазы информационной системы, интегрирующей источники данных информационных систем по свойствам неорганических веществ.

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 поле Status (тип int), которое используется для хранения статуса глобального идентификатора, который обсуждался в разделе 2.6. 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 [142] формате – это фактически множество химических элементов (рис. 3.8). Можно сказать, что таблица Meta_SystemInfo задает множество S, описанное в разделе 2.6, 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. 3.8. 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 разделе 2.6 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 [142] формате (рис. 3.9).

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

Рис. 3.9. 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 2.6. 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. 3.10).

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

Рис. 3.10. 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 2.6.

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.

3.2.2. 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:

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

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.

3.2.3. 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 2.6.3). 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 [150], 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. 3.11. 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 2.6.2).

```
  <?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. 3.11. 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. 3.12. 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. 3.12. Пример XML-документа, описывающего свойства, освещенные в информационном источнике (снимок экрана из Microsoft IE).

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. 3.13. 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).


```
  <?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. 3.13. 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.

3.2.4. 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 [151]. 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. 3.14.

```
  <?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. 3.14. 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. 3.14, приведен на рис. 3.15.

```
  <?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. 3.15. Пример XML-документа, передаваемого в качестве ответа Web-сервиса на запрос глобальных идентификаторов (снимок экрана из Microsoft IE).

Следует отметить, что данный XML-документ соответствует документу-запросу на рис. 3.14 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. Необходимо отметить, что в узлах документа-запроса, соответствующих путям “/MetaInfo/ChemicalSystemList/ChemicalSystem” и “/MetaInfo/ChemicalSubstanceList/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. 3.16.

```
  <?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. 3.16. 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. 3.16. 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. 3.16 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. Экспертный рейтинг добавляется к соответствующим узлам результирующего документа в качестве значения атрибута Rating. Узел DescribeChemicalEntities 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. 3.17.

```
  <?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. 3.17. Фрагмент XML-документа, содержащего ответ предметного посредника интегрируемой ИС на запрос значений свойств (снимок экрана из Microsoft IE).

Как видно из XML-документа на рис. 3.17, 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 третьей главе получены следующие результаты:

- разработана структура метабазы информационной системы, интегрирующей Web-приложения информационных систем по свойствам неорганических веществ;
- разработан и программно реализован механизм загрузки информации в метабазу информационной системы, интегрирующей Web-приложения;
- разработан и программно реализован механизм поиска релевантной информации в рамках интегрированной информационной системы;

- разработан и программно реализован механизм безопасного перехода пользователя между гетерогенными Web-приложениями в рамках контекста интегрированной информационной системы;
- разработана структура метабазы информационной системы, интегрирующей источники данных информационных систем по свойствам неорганических веществ;
- разработан и программно реализован алгоритм расчета достоверности информации, основанного на экспертных оценках;
- разработаны требования к реализации программных адаптеров интегрированной информационной системы;
- разработан и программно реализован предметный посредник интегрированной информационной системы.

ГЛАВА 4. 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 [152, 153]. Дальнейшая интеллектуализация баз данных непосредственно связана с оснащением информационных систем программными комплексами анализа огромных массивов химической информации, которые содержит разработанная в настоящей диссертационной 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 [154]. 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 закономерностей в химической информации. Найденные закономерности применены для конструирования новых неорганических соединений, перспективных для поиска новых соединений для электроники [155]. Целью проведенных исследований было выяснение возможностей создания информационно-аналитической системы для компьютерного конструирования неорганических веществ. Физико-химической основой разработки такой системы является Периодический закон, из которого следует, 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.

4.1. Этапы компьютерного конструирования новых соединений

Задача конструирования новых неорганических соединений сформулирована следующим образом [1, 133]: 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. 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, 133] с использованием методов обучения ЭВМ распознаванию образов [156-158]. Найденные закономерности могут быть представлены в виде ассоциативной структуры данных, например, искусственной нейронной сети [158] или растущей пирамидальной сети [156], а также в форме булевского выражения, продукционных правил [156, 158], системы алгебраических уравнений [157, 158] 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.

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 её 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 [156] или [158] 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 схема использования интегрированной ИС в качестве источника данных исследовательских систем, используемых для принятия решений при исследовании неорганических соединений (рис. 4.1).

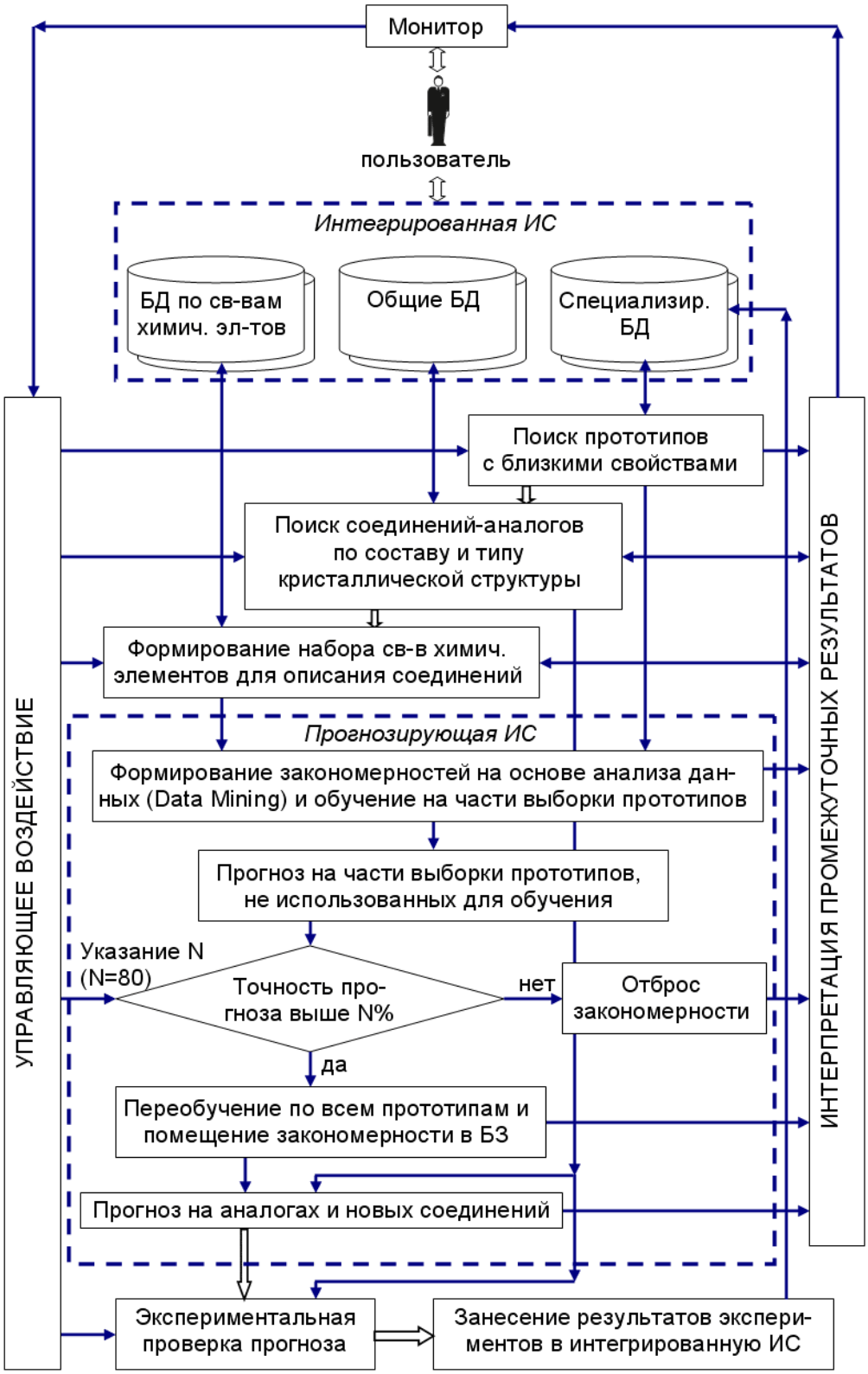

Рис. 4.1. Схема использования интегрированной ИС в СППР.

Прогнозирующая ИС используется, если в интегрированной ИС нет данных по соединениям с нужными свойствами. После экспериментальной проверки результатов прогноза, информация помещается в разработанную 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.

4.2. Перспективные полупроводники

Еще полвека назад академик Иоффе отметил в своей монографии [159], 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 [156] 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 таблице 4.1 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; “O” – 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 3 оказались неправильными.

Таблица 4.1. Часть таблицы прогноза 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	+	+	+	+	+	+	+	+	+		+	+	+	+	+	+	+	+

Далее был осуществлен прогноз типа кристаллической структуры при нормальном давлении и комнатной температуре для соединений, предсказанных выше [1, 133]. Для соединений состава ABX_2 с кристаллической структурой халькопирита была решена задача прогноза ширины запрещенной зоны с использованием программных комплексов обучения ЭВМ [156, 158]. Информация для компьютерного анализа о ширине запрещенной зоны известных халькопиритов этого состава была получена из БД “BandGap”. 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 таблице 4.2 даны результаты сравнения прогнозируемых значений ширины запрещенной зоны с экспериментальными значениями для различных алгоритмов, программы которых включены в системы [156, 158]. Приняты следующие обозначения: класс соединений: “**1**” – $E_g \leq 2$ эВ; класс “**2**” – $E_g > 2$ эВ; алгоритмы: “**I**” – алгоритмы вычисления оценок; “**II**” – линейный дискриминант Фишера; “**III**” – линейная машина; “**IV**” – логические закономерности; “**V**” – многослойный перцептрон; “**VI**” – Q ближайших соседей; “**VII**” – метод опорных векторов; “**VIII**” – статистически взвешенные синдромы; “**IX**” – голосование по тупиковым тестам [155]; “**X**” – Confor [156].

В таблице 4.3 приведена точность прогноза ширины запрещенной зоны известных халькопиритов с использованием вышеуказанных программ обучения ЭВМ [156, 158]. 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.

Таблица 4.2. 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 4.3. 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 перцептрон	87.1
Q ближайших соседей	77.4
метод опорных векторов	100
статистически взвешенные синдромы	80.6
голосование по тупиковым тестам	77.4
Confor	100

На основе анализа результатов табл. 4.2 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 [158]. В случае решаемой задачи это позволило получить 100%-ную точность прогноза.

4.3. Перспективные диэлектрики

Известно, что линейный электрооптический эффект, пьезоэлектрический эффект и эффект генерации второй гармоники возможны только в нецентросимметричных (ацентричных) кристаллах. Рез и Поплавко [160], 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 [156].

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 [160-163]. Таблица 4.4 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 пять из 17 проверенных результатов не совпали с экспериментом.

Таблица 4.4. 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					*	(*)	*	*©	*	*	(*)	*©	(*)	*¢	*	(*)	*	(*)	(*)	*

Таким образом, интеграция информационных ресурсов позволяет провести системный анализ больших массивов хранящейся информации с целью поиска

закономерностей в данных и использования их для конструирования новых веществ с заданными свойствами. Именно системный подход как комплексное взаимосвязанное последовательное рассмотрение всех факторов, путей и методов решения при поиске сложных закономерностей в данных должен быть основой для создания проектируемой информационно-аналитической системы для компьютерного конструирования неорганических веществ. В качестве информационной базы в нее должна входить разработанная нами интегрированная ИС, а в качестве инструментальных средств анализа данных могут использоваться методы искусственного интеллекта – обучение ЭВМ распознаванию образов.

Краткие выводы

В четвертой главе проведены эксперименты по использованию интегрированной ИС для анализа больших массивов хранящейся информации с целью поиска закономерностей в данных и использования их для конструирования новых веществ с заданными свойствами. Получены следующие результаты:

- Осуществлено конструирование еще не полученных перспективных полупроводниковых соединений состава ABX_2 (X = S, Se, Te). Сравнение результатов расчетов с новыми экспериментальными данными показало, что из 61 проверенного прогноза только 3 оказались неправильными, т.е. ошибка прогнозирования менее 5 %.
- Для соединений состава ABX_2 (X = S, Se, Te) с кристаллической структурой халькопирита решена задача прогноза ширины запрещенной зоны разными алгоритмами, при этом ошибка прогнозирования оказалась порядка 20 %, а при применении метода “голосования” результатов прогнозов с использованием различных алгоритмов удалось добиться правильных прогнозов.
- Осуществлен прогноз более сложных по составу соединений $A_2B_2(XO_4)_3$ с кристаллической структурой лангбейнита, перспективных для поиска новых пьезоэлектрических, сегнетоэлектрических, нелинейнооптических, электрооптических и люминесцентных веществ. Из 17 экспериментально проверенных прогнозов 12 оказались верными, т.е. ошибка прогнозирования менее 30 %.

На основе анализа полученных результатов сделан вывод о перспективности использования разработанной интегрированной информационной системы как информационной основы для программного комплекса компьютерного конструирования неорганических соединений.

ОСНОВНЫЕ РЕЗУЛЬТАТЫ И ВЫВОДЫ

В диссертационной работе получены следующие результаты:

- рассмотрены наиболее значимые базы данных по свойствам веществ и технологиям их получения, созданные в мире, и тенденции их развития;
- рассмотрены информационные системы ИМЕТ РАН, построенные на разных программно-аппаратных платформах и с использованием разных подходов к хранению и обработке информации;
- проанализированы современные технологии интеграции гетерогенных информационных систем, выработаны критерии и осуществлен выбор программной платформы для реализации информационной системы;
- на основе проведенного анализа архитектуры информационных систем по свойствам веществ и технологиям их получения разработан комплексный подход к интеграции информационных систем, как на уровне пользовательских интерфейсов, так и на уровне информационных источников;
- на основе теории множеств формализовано понятие релевантной информации, содержащейся в интегрируемых информационных системах рассматриваемой предметной области;
- разработана методика интеграции распределенных Web-приложений информационных систем с учетом требований информационной безопасности, оснащенная поисковым механизмом для обнаружения релевантной информации, которая содержится в интегрируемых информационных системах;
- разработана методика интеграции разнородных информационных источников на основе предложенных механизмов разрешения конфликтов гетерогенности и схемы предметной области;
- разработан, реализован и внедрен в ИМЕТ РАН программный комплекс, интегрирующий информационные системы по свойствам веществ и технологиям их получения, который позволяет пользователям осуществлять доступ ко всей информации, хранящейся в интегрированных информационных системах;
- использование интегрированной ИС для анализа больших массивов хранящейся информации с целью поиска взаимосвязей в данных и использования их для прогнозирования свойств веществ показало, что средняя ошибка прогнозирования – менее 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. 10 years STN International. Databases in Science & Technology. 1993. FIZ Karlsruhe. STN Service Center Europe. – 45 p.
18. 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.
19. Ларичев О.И., Мошкович Е.М. Качественные методы принятия решений. – М., Наука. Физматлит. 1996.
20. Кафаров В.В., Дорохов И.Н. Системный анализ процессов химической технологии. – М., Наука, 1976, 500 с.

21. Волкова В.Н., Денисов А.А. Основы теории систем и системного анализа. – СПб: СПбГТУ, 2001, 512 с.
22. Dr. Brodie, M. The Grand Challenge of Information Technology - Invited talk, CAiSE-2002.
23. http://www.forresterresearch.com
24. Lopatenko A. National networks of science and technology information. Semantic Web based architecture for access to research and technology data – Proc. of EVA2001, Moscow.
25. Imhoff C. Understanding the Three E's of Integration EAI, EII and ETL. Intelligent Solutions, Inc, April 2005. Available at http://www.intelsols.com.
26. Halevy A. et all. Enterprise Information Integration: Successes, Challenges and Controversies – SIGACM-SIGMOD 2005 Baltimore, Maryland USA
27. http://www.w3.org/XML/Query/
28. Inmon W.H. Building the Data Warehouse. – New-York, John Wiley, 1992.
29. Bitton D. Why EII will not replace the data warehouse – SIGACM-SIGMOD 2005 Baltimore, Maryland USA
30. http://www.tpc.org
31. Draper D. The Nimble experience – SIGACM-SIGMOD 2005 Baltimore, Maryland USA
32. 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 17 (1999). pp. 270-293.
33. Bergamaschi S., Castano S., Vincini M., Beneventano D.: Semantic integration of heterogeneous information sources. – Data and Knowledge Engineering 36 (2001). pp. 215-249.
34. Cali A., Calvanese D., Giacomo G.D., Lenzerini M., Naggar P., Vernacotola F.: Ibis: Semantic data integration at work. – Lecture Notes in Computer Science 2681 (2003). pp. 79-94.
35. Levy A.: Logic-based techniques in data integration. In Minker J., ed.: Logic Based Artificial Intelligence, Kluwer Academic Publishers (2000).
36. Halevy A.Y.: Answering queries using views. Very Large Databases 10 (2001). pp. 270-294.
37. Lenzerini M.: Data integration: A theoretical perspective. In: 21st ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems. (2002). pp. 233-246.
38. 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). pp. 241-250.
39. Convent B.: Unsolvable problems related to the view integration approach. Proc. of International Conference on Database Theory (1986) 141-156.

40. Lopatenko A. Query answering under Exact View Assumption in Local As View Data Integration System. Proc. of EVA2001, Moscow.
41. Bravo L., and Bertossi L. Disjunctive deductive databases for computing certain and consistent answers to queries from mediated data integration systems. Journal of Applied Logic, 2004.
42. Imielienski T., Witold Lipski J. Incomplete information in relational databases. Journal of the ACM (JACM) 1984. 31(4):761-791.
43. Kolaitis P. Course on constraint satisfaction, complexity, and logic. In ESSLLI, Vienna, Austria, 2003.
44. 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.
45. http://www.crystalimpact.com/pauling.
46. 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.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. Киселева Н.Н., Прокошев И.В., Дударев В.А. и др. Система баз данных по материалам для электроники в сети Интернет // Неорганические материалы. 2004. Т.42. №3. - С.380-384.
54. https://webrech.fiz-karlsruhe.de/CATALOG/newcryst.html.
55. Tomaszewski P.E. Structural phase transitions in crystals. I. Database // Phase Transit. 1992. V.38. №3. - P.127-220.
56. http://www.codata.org/databases/Materials.html.

57. Faber J., Fawcett T. The Powder Diffraction File: present and future // Acta Crystallogr. 2002. V.B58. №3. - P.325-332.
58. http://www.cryst.ehu.es/icsdb/about.html.
59. 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.
60. 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.
61. 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.
62. Вертопрахов В.Н., Доленко Т.Н., Кучумов Б.М. Фактография в электронном материаловедении. Новосибирск: Наука, 1988. - 101 с.
63. 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.
64. Гурвич Л.В. ИВТАНТЕРМО – автоматизированная система данных о термодинамических свойствах веществ // Вестн.АН СССР. 1983. №3. - С.54-65.
65. 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.
66. Cheynet B. THERMODATA: an integrated thermodynamic and inorganic physico-chemical information system // CODATA Bull. 1985. №58. - P.18-22.
67. https://webrech.fiz-karlsruhe.de/webrech/DBSS/trcthermoss.html.
68. http://ultra.ippe.obninsk.ru:8097/nea_databank/dbsurvey.htm.
69. http://www.cas.org/ONLINE/DBSS/jicsteplusss.html.
70. Westhaus U., Droge T., Sass R. DETHERM – a thermophysical property database // Fluid Phase Equil. 1999. №158-160. - P.429-435.
71. 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.
72. Palmer D.A. DIPPR – an improbable success // AIChE Symp.Ser. 1990. V.86. №275. - P.1-4.
73. http://www.nist.gov/srd/nist17.htm.
74. http://properties.nist.gov/fluidsci/semiprop/.
75. 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.
76. https://cindasdata.com/.

77. Трусов Б.Г., Стрельцов Ф.Н., Огнивов В.В. Использование автоматизированной системы термодинамических расчетов в технологических исследованиях // III Всес.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.
78. 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.
79. Dinsdale A.T. SGTE data for pure elements // CALPHAD. 1991. V.15. №4. - P.317-425.
80. 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.
81. 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.
82. 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.
83. http://www.kbk-sdi.com/sdisp/nel/ppds_window/PPDS_Window_Iss6.pdf.
84. Asada Y., Nakada E., Yokokawa T., et al. Database for materials design of multi-layered superconductors // J.Mater.Sci.Soc.Jap. 1988. V.24. №4. - P.199-203.
85. Chen H., Iwata S. Data analysis by a data system on high-Tc superconducting materials // Materials System. 1993. V.12. Nov. - P.63-70.
86. http://www.cas.org/ONLINE/DBSS/asmdatass.html.
87. http://www.cas.org/ONLINE/DBSS/copperdatass.html.
88. http://www.bus.iastate.edu/mennecke/server/courses/RICDBSpecifications.htm
89. 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.
90. http://theorie.physik.uni-wuerzburg.de/webrech/DBSS/mdfss.html.
91. 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.
92. 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.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.
98. 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.
99. 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.
100. Japan’s first move towards collaborative ventures // III-Vs Rev. The Adv.Semicond.Mag. 2003. V.16. №8. - P.6.
101. Ansara I. Generation et application des bases de donnees thermochimiques // Entropie. 1991. V.27. №161. - P.74-79.
102. Липаев В.В. Технико-экономическое обоснование проектов сложных программных средств. М.: Синтег, 2004. - 284 с.
103. Microsoft Windows Server 2003 with Internet Information Services (IIS) 6.0 vs. Linux Competitive Web Server Performance Comparison – Veritest report. April, 2003. Available at http://www.veritest.com.
104. Koetzle L., Rutstein Ch., Lambert N., Wenninger S. Is Linux more secure than Windows? – Forrester Research Inc. March 19, 2004. Available at http://www.forrester.com.
105. Смагин В.А., Солдатенко В.С., Кузнецов В.В. Моделирование и обеспечение надежности программных средств АСУ. – СПб. 1999. 49 с.
106. Microsoft Windows Server 2003 vs. Red Hat Enterprise Linux AS 3.0: IT Professionals Running a Production Environment – Veritest report. April, 2005. Available at http://www.veritest.com.
107. EICTA Interoperability White Paper – June 21, 2004. Available at http://www.eicta.org.
108. http://www.w3.org/2002/ws/
109. Wilcox J., Sargent P., Bayriamova Z., Matiesanu C. Interoperability: How Technology Managers Rate Microsoft and Its Technologies for Development – Jupiter Research (MIC04-C02). April 7, 2004. Available at http://www.jupiterresearch.com.

110. Roger Sessions. Interoperability Through Service-Oriented Architectures (SOAs). ObjectWatch. Available at http://www.objectwatch.com.
111. Magic Quadrant for WS Major Vendor Influence, 3Q03 – Gartner Research. Available at http://www.gartner.com.
112. Rymer J., Cormier B. The Total Economic Impact of Developing and Deploying Applications on Microsoft and J2EE/Linux Platforms – September 4, 2003. Available at http://www.forrester.com.
113. Server operating system licensing & support cost comparison Platforms – May 2004. Available at http://www.bearingpoint.com.
114. http://www.tpc.org
115. 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 – стр. 18-19.
116. Дударев В.А. Универсальная программа удаленного администрирования баз данных. “Научный сервис в сети Интернет”. Труды Всероссийской научной конференции. М.: МГУ, 2002 – стр. 75-77.
117. Миано Дж. Форматы и алгоритмы сжатия изображений в действии. Изд.: Триумф, 2003. – 336 с.
118. Дударев В.А. Программа удаленного администрирования базы данных по свойствам кристаллов акустооптических, электрооптических и нелинейнооптических веществ. “Новые информационные технологии”. Тезисы докладов. Том 2. М.: МГИЭМ, 2002 – стр. 359-360.
119. Wilson A.H. The Theory of Metals, Cambridge, 1953.
120. Regel A.R., Glazov B.M. Periodical Law and Physical Properties of Electronic Melts, Moscow, 1978 (in Russian).
121. Косяков В.И., Сурков Н.В. Способы обработки и хранения информации о фазовых диаграммах // Геология и геофизика. 1998. Т.39. №9. - С.1192-1209.
122. Масленков С.Б., Удовский А.Л. Банки данных по диаграммам состояний металлических систем: современное состояние, проблемы, перспективы их развития и применения // В сб. Расчеты и экспериментальные методы построения диаграмм состояния. М.: Наука, 1985. - С.77-87.
123. Nash P. Computer representation of phase diagrams // Bull. Alloy Phase Diagr. 1984. V.5. №1. - P.5-9.
124. Кауфман Л., Бернстейн Х. Расчет диаграмм состояния с помощью ЭВМ. М.: Мир, 1972. - 326 с.
125. Новик Ф.С., Кожевников И.Ю., Гультяй И.И. Банки данных по диаграммам состояния в автоматизированных системах научных исследований // В сб. Расчеты и экспериментальные методы построения диаграмм состояния. М.: Наука, 1985. - С.87-93.

126. Murrey I., Orser I. Interactive computer graphics for storing of phase diagram // Bull. Alloy Phase Diagr. 1980. V.1. №1. - P.19-31.
127. Дегтярев С.А., Воронин Г.Ф. Применение сплайнов в термодинамике растворов // В сб. Математические проблемы фазовых равновесий. Новосибирск: Наука, 1983. - С.53-83.
128. Косяков В.И., Малахов Д.В. Принципы свертки и хранения информации о фазовых диаграммах // В сб. Прямые и обратные задачи химической термодинамики. Новосибирск: Наука, 1987. - С.73-80.
129. Луцык В.И., Воробьева В.П., Сумкина О.Г. Моделирование фазовых диаграмм четверных систем. Новосибирск: Наука, 1992. – 199 с.
130. Armstrong J., deHaan J. Macromedia Flash 8. Фирменное руководство. Изд.: Триумф, 2006. – 256 с.
131. Паркер Т., Сиян К. TCP/IP для профессионалов. – Изд.: Питер, 2004. 864 с.
132. Кришнамурти Б., Рексфорд Дж. Web-протоколы. Теория и практика. HTTP/1.1, взаимодействие протоколов, кэширование, измерение трафика. - Изд.: Бином, 2002. 592 с.
133. Киселева Н.Н. Компьютерное конструирование неорганических соединений, перспективных для применения в электронике, с использованием баз данных и методов искусственного интеллекта: Дисс. докт. хим. наук. Москва, МИТХТ. 2004. – 333 с.
134. Дударев В.А. Подходы к интеграции российских баз данных по материалам для электроники. “Новые информационные технологии”. Тезисы докладов. М.: МГИЭМ, 2004 – стр. 83-85.
135. Дударев В.А. Создание интегрированной системы баз данных по материаловедению. Труды молодых специалистов ИМЕТ им. А.А. Байкова РАН 2005: Сб. статей под ред. Ю.К. Ковнеристого и др. – Москва, Издательство Интерконтакт Наука, 2005 – стр. 148-150.
136. Kornyshko V., Dudarev V. Software Development for Distributed System of Russian Databases on Electronics Materials // Int. Journal “Information Theories & Applications”, vol. 13, number 2, 2006. pp. 121-126.
137. Дударев В.А. Разработка программного обеспечения интегрированной системы баз данных ИМЕТ РАН по свойствам материалов для электроники. “Научный сервис в сети Интернет: технологии распределенных вычислений”. Труды Всероссийской научной конференции. М.: МГУ, 2005 – стр. 227-229.
138. Полянский А. Программирование на CGI. Изд. Майор, 2003 – 176 с.
139. Валиков А. Технология XSLT. БХВ-Петербург, 2002 – 544 с.
140. Ньюкомер Э. Веб-сервисы: XML, WSDL, SOAP и UDDI. Изд.: Питер, 2003. – 256 с.
141. Орлик С.В. Обзор технологических стандартов Web-служб и тенденций их развития. // Сборник трудов II-й Всероссийской практической конференции "Стандарты в проектах современных информационных систем", М., 2002 г.

142. http://www.w3.org/XML/Schema
143. Чаусов В.И. Практическое применение языка XML в задаче интеграции бизнес-приложений. // Сборник трудов II-й Всероссийской практической конференции "Стандарты в проектах современных информационных систем", М., 2002 г.
144. Дударев В.А. Реализация Интегрированной информационной системы, объединяющей Web-интерфейсы информационных систем по свойствам неорганических веществ и материалов. Руководство разработчика. Доступно по адресу http://meta.imet-db.ru/LinkIntegrationManual.doc
145. Дударев В.А. XML-schema стандартизирующая формат XML-документа для обновления метабазы доступна по адресу http://meta.imet-db.ru/MUService.xsd.
146. Дударев В.А. WSDL-контракт, оговаривающий методы взаимодействия с Web-сервисом обновления метабазы, доступен по адресу http://meta.imet-db.ru/MUService/MUService.asmx?wsdl.
147. 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).
148. Дударев В.А. WSDL-контракт, оговаривающий методы взаимодействия с Web-сервисом обслуживающим интегрируемые ресурсы, доступен по адресу http://meta.imet-db.ru/Service/Service.asmx?wsdl.
149. Дударев В.А. XML-schema стандартизирующая формат XML-документа, возвращаемого Web-сервисом поиска релевантной информации, со списком релевантной информации доступна по адресу http://meta.imet-db.ru/Relevance.xsd.
150. Дударев В.А. WSDL-контракт, оговаривающий методы взаимодействия с программными адаптерами интегрируемых информационных систем выполненными в виде Web-сервисов, доступен по адресу https://crystal.imet-db.ru/EII_Crystal/EII_Crystal.asmx?WSDL.
151. Дударев В.А. WSDL-контракт, оговаривающий методы взаимодействия с предметным посредником интегрированной информационной системы, доступен по адресу https://meta.imet-db.ru/EII/Service.asmx?WSDL.
152. Kiselyova N., Dudarev V. The Distributed System of Databases on Properties of Inorganic Substances and Materials. // Int. Journal “Information Theories & Applications”, vol. 12, number 3, 2005. pp. 219-224.
153. Дударев В.А., Киселева Н.Н., Земсков В.С. Интегрированная система баз данных по свойствам материалов для электроники // Перспективные материалы, 2006, №5 – стр. 20-25.

154. Дударев В.А., Колыбанов К.Ю. Повышение экономической эффективности разработки полифункциональных материалов на основе интеграции баз данных. Ученые записки МИТХТ, выпуск 14: Сб. статей под ред. В.С. Тимофеева и др. – М.: МИТХТ, 2005 – стр. 79-80.
155. 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 – стр. 61-68.
156. Гладун В.П. Процессы формирования новых знаний. София: СД "Педагог 6", 1995. – 192 с.
157. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: изд-во Ин-та математики, 1999. – 269 с.
158. Журавлев Ю.И., Рязанов В.В., Сенько О.В. «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. М.:изд-во ФАЗИС, 2006. – 176 с.
159. Иоффе А.Ф. Физика полупроводников. М.-Л.: изд-во АН СССР, 1957. – 491 с.
160. Рез И.С., Поплавко Ю.М. Диэлектрики. Основные свойства и применения в электронике. М.: Радио и связь, 1989. – 288 с.
161. Ярив А., Юх П. Оптические волны в кристаллах. М.: Мир, 1987. – 616 с.
162. 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.
163. 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.

ПРИЛОЖЕНИЕ