УДК 004.41
DOI: 10.15827/0236-235X.130.283-296
Дата подачи статьи: 30.01.20
2020. Т. 33. № 2. С. 283–296

Анализ методов интеграции для разработки информационно-аналитических систем по свойствам неорганических соединений

В.А. Дударев [1], *к.т.н., доцент, vdudarev@hse.ru*
И.О. Темкин [2], *д.т.н., профессор, igortemkin@yandex.ru*
В.Ф. Корнюшко [3], *д.т.н., профессор, vfk256@mail.ru*

[1] *Национальный исследовательский университет «Высшая школа экономики», г. Москва, 109028, Россия*
[2] *Национальный исследовательский технологический университет «МИСиС», г. Москва, 119049, Россия*
[3] *МИРЭА – Российский технологический университет, г. Москва, 119454, Россия*

В статье рассматривается применение системного анализа для формализации информационных процессов, протекающих в интегрируемых информационных системах, и разработки системы рекомендаций для выбора одного из методов интеграции при решении практических задач консолидации информационных систем.

Анализ начинается с выделения базовых информационных процессов, протекающих в локальных информационных системах, затем идет наблюдение за их трансформацией в интегрированных системах, построенных с использованием трех методов консолидации: интеграции корпоративных приложений (Enterprise Application Integration, EAI), интеграции корпоративной информации (Enterprise Information Integration, EII) и ПО для извлечения, преобразования и загрузки данных (Extract, Transform, Load – ETL), основанного на технологии хранилищ данных.

Осуществлен синтез обобщенной схемы методов интеграции гетерогенных информационных систем, и на ее основе выполнено сравнение методов с использованием десяти критериев. По итогам сравнения предложены рекомендации по выбору предпочтительного метода интеграции гетерогенных информационных систем.

Отмечается возможность совместного использования методов ETL и EII для достижения оптимальной скорости работы источника интегрированных данных. Поскольку ни один из методов не покрывает всех потребностей, возникающих при консолидации информационных систем, предлагается методология их совместного использования для наиболее тесной интеграции информационных систем.

Кратко рассматриваются результаты практического применения разработанной методологии интеграции для консолидации информационных систем в области неорганического материаловедения, как отечественных, так и зарубежных: описывается набор созданных web-сервисов, являющихся основным техническим средством, обеспечивающим взаимодействие распределенных кроссплатформенных информационных систем по свойствам неорганических веществ для консолидации данных, и пользовательский интерфейс единой точки входа www.imet-db.ru, позволяющий любому пользователю осуществлять поиск и просмотр информации по свойствам содержащихся в интегрируемых информационных системах веществ.

Ключевые слова*: информационный процесс, интеграция данных, интеграция приложений, сравнение методов интеграции, критерий выбора метода интеграции.*

Задача интеграции информации в настоящее время актуальна для многих организаций, поскольку позволяет повысить эффективность работы. Этим объясняются большой интерес к данному направлению развития информационных технологий и появление множества новых программных продуктов, направленных на решение задач интеграции. Однако зачастую компании по-разному понимают интеграцию и, следовательно, по-разному подходят к решению задач интеграции. Это происходит на фоне еще не вполне четко сформировавшегося, размытого терминологического аппарата. Необходимость разъяснения сути методов интеграции и их преимуществ привела в июле 2001 года к созданию лидерами в области интеграции международного консорциума по интеграции (Integration Consortium – IC). Следует отметить, что до мая 2004 года у консорциума IC было другое название – консорциум отрасли интеграции корпоративных приложений (EAI Industry Consortium – EAIIC), которое изме-

нили, поскольку консорциум занимался всеми вопросами интеграции, а EAI является лишь одним из методов интеграции. В настоящее время IC – это международная некоммерческая организация, объединяющая более 50 компаний из различных стран мира. В работе IC принимают участие не только поставщики программного и аппаратного обеспечения и системные интеграторы, но и потребители методов интеграции, представители научных кругов. Поскольку IC задумывался как сообщество, целью которого является единение отрасли интеграции, все члены консорциума могут совместно определять проблемы и разрабатывать решения. По сути роль консорциума IC в сфере интеграции эквивалентна роли консорциума W3C в области web-технологий.

В данной работе авторы придерживаются термина «метод интеграции» вместо «технологии интеграции», так как его использование более уместно при разработке методологии интеграции *информационных систем* (ИС) по свойствам неорганических веществ, используемой для консолидации информационных ресурсов в области неорганического материаловедения. Методология рассматривается как система методов исследований, в данной работе – методов (или технологий) интеграции. А метод является набором методик, то есть совокупностью приемов практической реализации.

В настоящее время происходит не только становление терминологической базы в области интеграции, но и развитие самих интеграционных подходов. Обычно выделяют три метода интеграции: интеграция корпоративных приложений (Enterprise Application Integration, EAI), интеграция корпоративной информации (Enterprise Information Integration, EII) и ПО для извлечения, преобразования и загрузки данных (Extract, Transform, Load – ETL). Вследствие этого наблюдается некоторая неоднозначность относительно того, каковы функции каждого из трех описанных методов и в каких случаях тот или иной метод должен использоваться. Необходимо четко представлять возможности каждого метода и определять класс задач, для решения которых он подходит. Для понимания различий в назначении методов интеграции необходимо привести соответствующие определения, учитывающие их назначение [1].

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

Для более полного понимания этих методов необходимо рассмотреть их взаимосвязь в рамках уже существующей информационной инфраструктуры организации.

Базовые информационные процессы в локальных ИС

Типовая структура ИС включает в себя ряд подсистем, реализующих базовые информационные процессы сбора, хранения, передачи, обработки и представления информации. На рисунке 1 представлены информационные процессы, протекающие в локальной ИС. В ней реализуются все основные процессы (кроме информационного обмена с внешними ИС).

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

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

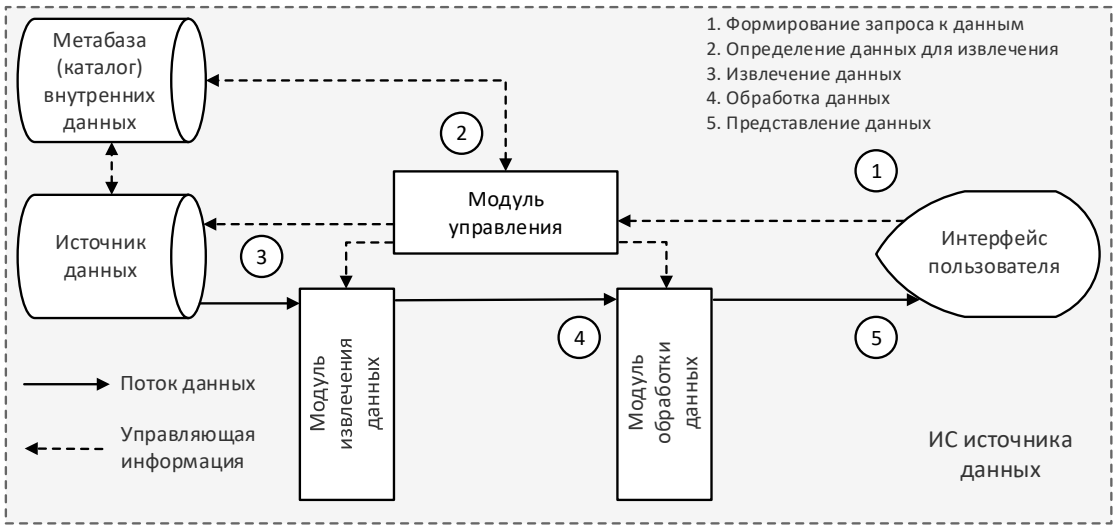

Рис. 1. Информационные процессы в локальной ИС

Fig. 1. Information processes in the local IS

Создание централизованной ИС, 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, способных объединить всю важную информацию, накопленную исследователями данной организации, актуальна при создании практически любой централизованной ИС.

При переходе к интегрированным ИС необходимо прежде всего ответить на следующие вопросы:

– какие подсистемы интегрированной ИС будут распределенными, а какие останутся локальными;

– какие подсистемы интегрированной ИС станут (изначально или в перспективе) гетерогенными, а какие останутся однородными;

– каков будет баланс между централизацией и периферийностью в системе управления интегрированной ИС.

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

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

Основной задачей при разработке централизованных систем является стандартизация, которой 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 ИС [2].

Метод интеграции корпоративной информации EII

Интеграция корпоративной информации – это интеграция данных из многочисленных систем в унифицированное, согласованное и точное представление, предназначенное для изучения и обработки данных [3].

При организации процесса интеграции данных по технологии EII главным функциональным модулем является предметный посредник (иногда называемый модулем извлечения), который обеспечивает единый интерфейс взаимодействия конечных приложений с источниками исходной информации, поиск запрашиваемой информации по исходным БД и агрегацию собранной информации для передачи конечным приложениям.

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

Схема интеграции разнородных источников данных на основе метода интеграции корпоративной информации показана на рисунке 2.

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

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

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

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

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

Метод интеграции на основе хранилищ данных ETL

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

Интеграция разнородных источников данных включает предварительное формирование хранилища данных и последующую работу с данными, размещенными не в ИС источников данных, а в хранилище данных [4].

Формирование хранилища данных осуществляется в три этапа.

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

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

При первичном извлечении данные берутся из БД источника в полном объеме. При после-

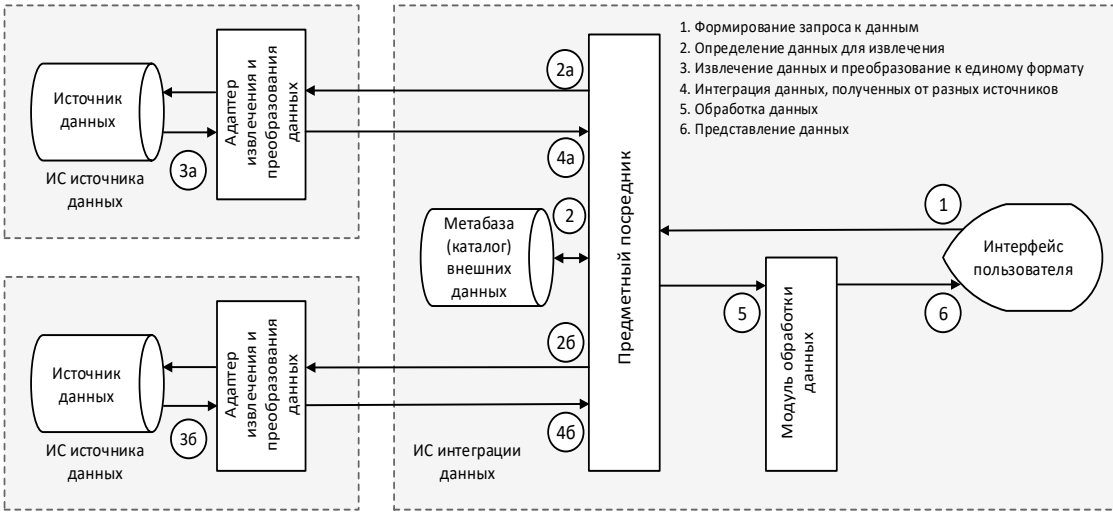

Рис. 2. Функционирование метода интеграции корпоративной информации EII

Fig. 2. EII corporate information integration method

дующих извлечениях для оптимизации работы системы может быть реализован механизм определения изменений данных источника и извлечения только тех, которые необходимы для актуализации информации в промежуточном хранилище (staging area).

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

На втором этапе с помощью функций преобразования осуществляется унификация представления данных промежуточных хранилищ для создания единой структуры хранения и организации данных. Выполняются объединение и слияние или, наоборот, разделение данных, изменение формата представления данных (например, реорганизация таблиц и отношений между таблицами), добавление новых атрибутов, сортировка и фильтрация. Также осуществляются анализ и контроль качества и полноты собранных данных, устраняются конфликты их интеграции.

По завершении данного этапа информация в промежуточных хранилищах приводится в единый формат, определяющий взаимодействие сформированной БД с инструментальными панелями и ПО.

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

После выполнения функций загрузки формируется база интегрированных данных, имеющая единую детерминированную структуру и интерфейс, с помощью которого любые модули и приложения могут обращаться к информации, хранящейся в базе.

Функции ETL могут одновременно применяться к нескольким БД источников либо к группам БД источников в случае их однотипности.

Схема интеграции разнородных источников данных на основе хранилища данных представлена на рисунке 3.

Последующая работа с хранилищем данных не отличается от работы с локальной БД, при этом обеспечивается высокая скорость работы с данными [5]. В ИС интеграции на основе хранилища данных реализуются все базовые информационные процессы (рис. 3) обработки информации.

Интеграция корпоративных приложений EAI

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

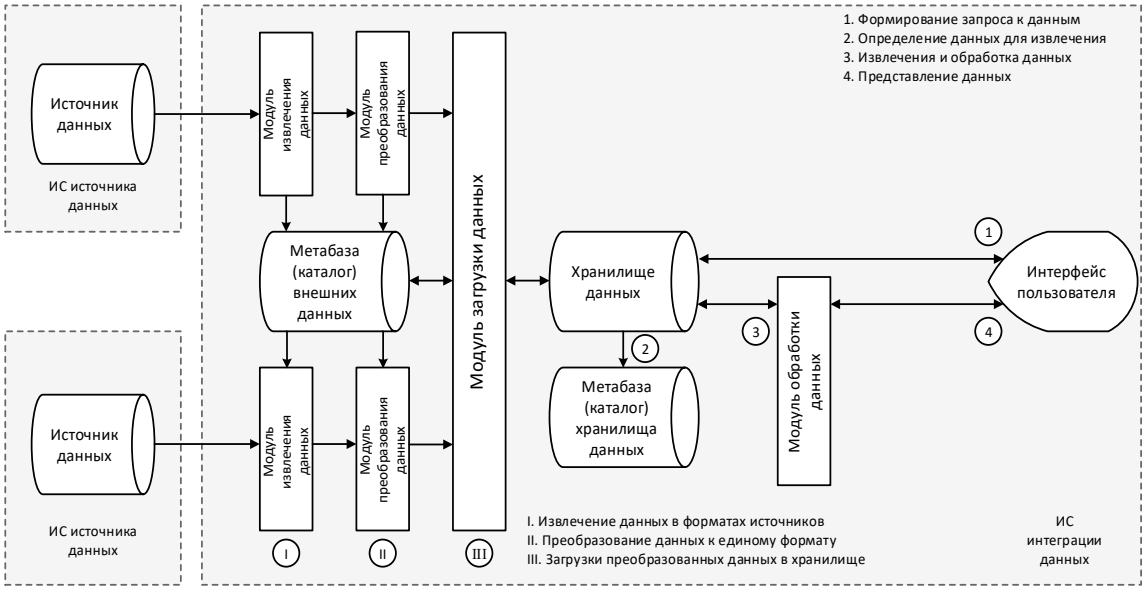

Рис. 3. Интеграция на основе хранилищ данных ETL

Fig. 3. Integration based on ETL data stores

Метод EAI позволяет автоматизировать процессы работы с разнородными данными без необходимости непосредственного обращения к данным и изменения готовых интерфейсов, программ и приложений работы с данными [6].

В контексте данного метода интеграции основной задачей является организация взаимодействия между объединенным интерфейсом работы с приложениями и приложениями-источниками – согласование формата, средств и способов передачи данных от одного приложения к другому.

Существуют несколько наиболее распространенных методов решения данной задачи:

– использование программных адаптеров (adapters) для обоих приложений;

– использование промежуточного ПО, ориентированного на обработку сообщений (Message-oriented middleware, MOM);

– использование репликатора данных (Data Replication Engine, DRE).

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

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

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

Схема интеграции разнородных источников данных на основе метода интеграции корпоративных приложений представлена на рисунке 4.

Иногда при использовании метода интеграции корпоративных приложений EAI дополнительно уточняется, какие именно корпоративные приложения имеются в виду – относящиеся к одной корпорации или к разным. В рамках одной организации интеграция корпоративных приложений обычно описывается термином Business Process Integration (BPI – интеграция бизнес-процессов). Если же речь идет об интеграции ИС разных организаций, то такую интеграцию часто называют B2B-интеграцией (Business-to-Business) [7].

Обобщенная схема методов интеграции гетерогенных ИС

Появление каждого из описанных выше методов интеграции обусловлено необходимо-

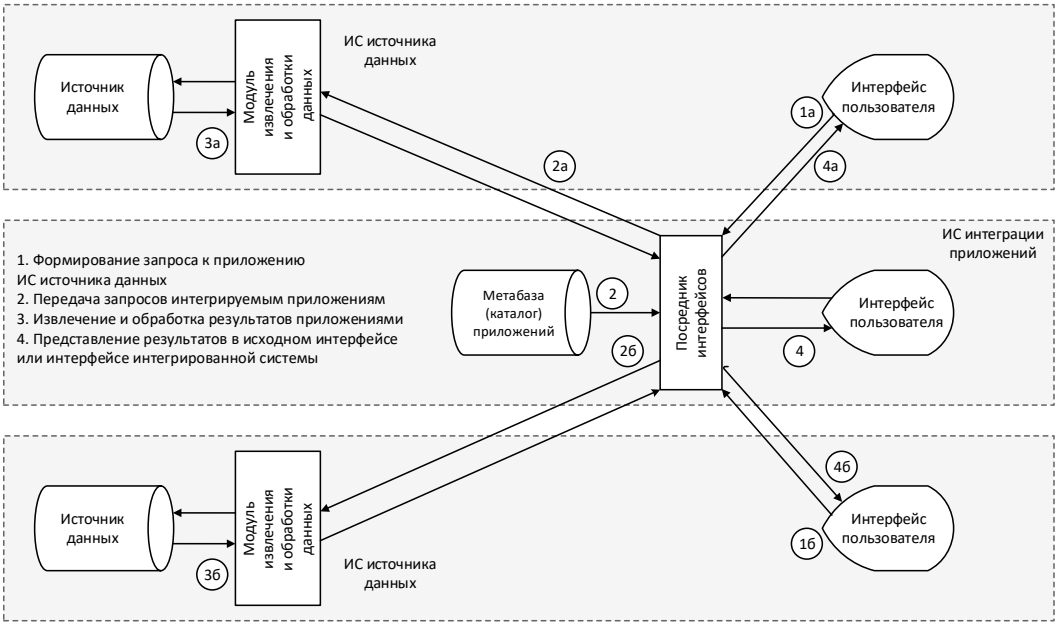

Рис. 4. Интеграция корпоративных приложений EAI

Fig. 4. Enterprise application integration

стью решения определенного круга задач, которые независимо от отрасли или характера деятельности возникали перед компаниями и организациями с ростом объемов используемых данных и расширением ИС.

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

В таблице 1 приведены критерии сравнения методов интеграции гетерогенных ИС для подбора наиболее подходящего варианта реализации интеграции для каждого конкретного случая.

При объединении ИС информационные процессы 1–5 (рис. 1) будут реализованы в различных ИС (множественных ИС источников данных либо в центральной ИС интеграции) при помощи специализированных программных компонентов (модулей). На основе системного анализа информационных потоков составлена обобщенная схема интеграции гетерогенных ИС (рис. 5) [8]. Пунктиром на схеме показаны условные границы интегрируемых ИС.

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

Интеграция на основе метода хранилищ данных ETL включает модули извлечения ис-

Таблица 1

Критерии сравнения методов интеграции гетерогенных ИС

Table 1

Criteria for comparing integration methods of heterogeneous IS

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

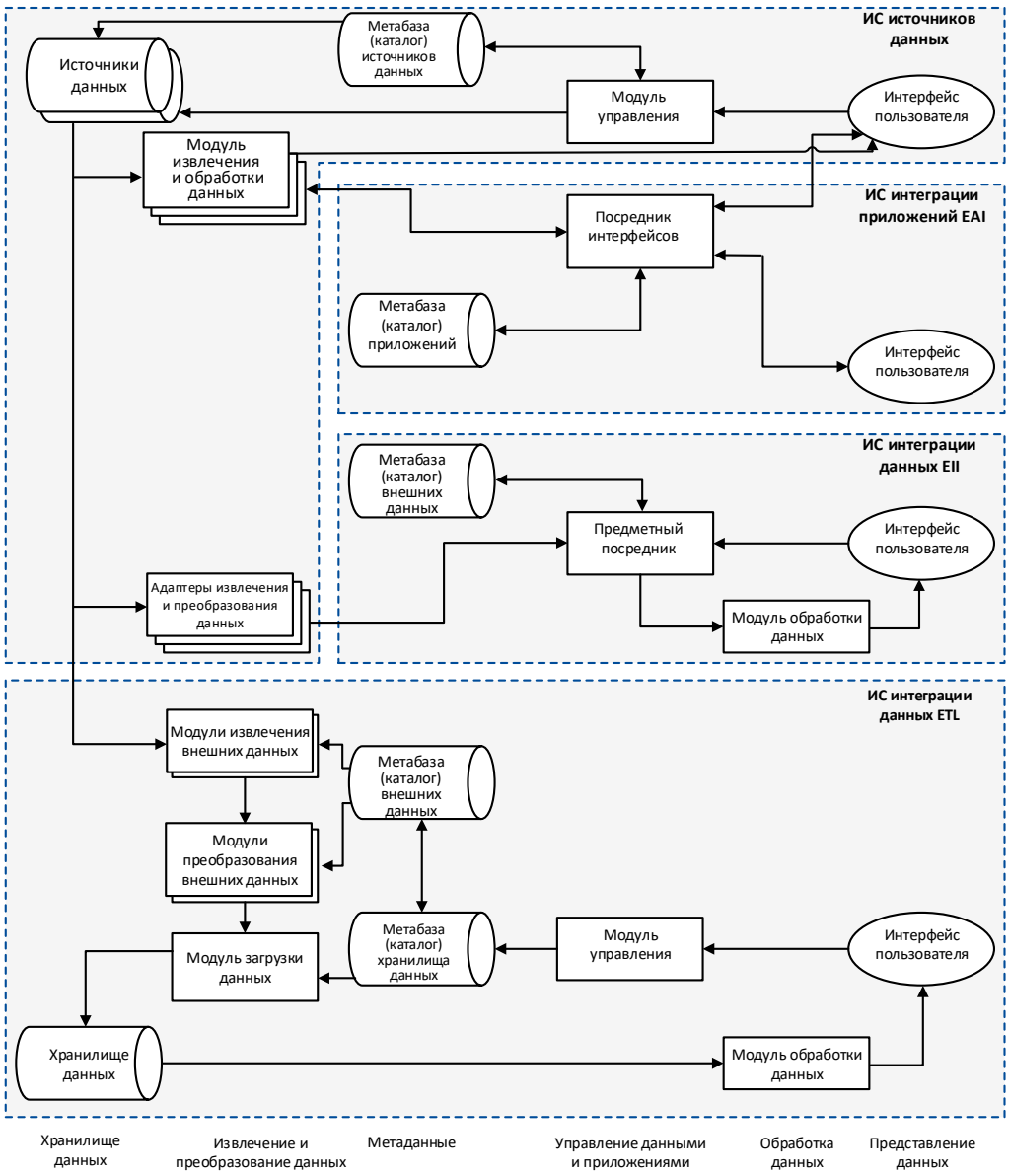

Рис. 5. Обобщенная схема интеграции гетерогенных ИС

Fig. 5. Integration generalized scheme of heterogeneous IS

ходных данных в форматах ИС источников (на основе метабазы внешних данных), преобразования их к формату хранилища данных и загрузки в локальное хранилище (на основе метабазы хранилища данных) [9]. Локальное расположение всех модулей обработки данных требует доступности ИС источников только на момент первичного извлечения данных.

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

При интеграции гетерогенных ИС (в отличие от локальной ИС) необходима реализация процессов внешнего информационного обмена. На обобщенной схеме интеграции (рис. 5) эти процессы показаны стрелками информационных потоков, пересекающими условные границы ИС. Также процессы передачи информации имеют место при реализации удаленного доступа пользователей к интерфейсу ИС интеграции.

Принципы интеграции, заложенные в рассмотренных методах, используются для реше-

ния широкого круга задач: от интеграции в режиме реального времени до пакетной интеграции и от интеграции данных до интеграции приложений. На рисунке 6 показано положение названных методов по отношению к этим двум типам задач. Для интеграции данных в режиме реального времени более рационален подход EII, для пакетной интеграции данных – ETL. Для интеграции приложений в режиме реального времени или пакетном наиболее подходит метод EAI.

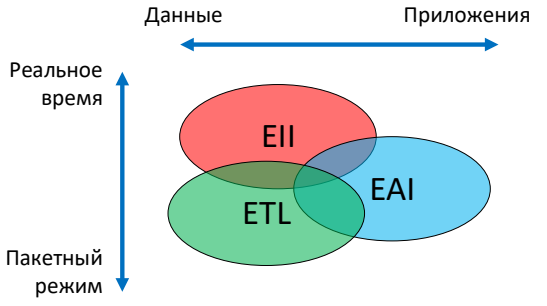

Рис. 6. Современные методы интеграции ИС

Fig. 6. IS modern integration methods

Рекомендации по выбору метода интеграции

В результате анализа критериев сравнения методов интеграции (табл. 1) и обобщенной схемы интеграции гетерогенных ИС (рис. 5) можно определить ряд ситуаций, в которых использование конкретного метода интеграции является предпочтительным либо единственно возможным. Рекомендации по выбору предпочтительного метода интеграции гетерогенных ИС приведены в таблице 2.

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

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

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

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

Невозможность полного доступа к данным ИС источника исключает применение метода

Таблица 2

Рекомендации по выбору предпочтительного метода интеграции гетерогенных ИС

Table 2

Recommendations for choosing the preferred method for integrating heterogeneous IS

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

хранилищ данных. Платный доступ к данным ИС источника определяет высокую стоимость хранилища данных и делает его разработку экономически неэффективной.

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

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

Консолидация ИС в области неорганического материаловедения

Следует отметить, что ни один из существующих на сегодняшний день методов интеграции не способен решить все проблемы, возникающие в области неорганического материаловедения при объединении ИС, разработанных в разных организациях и странах [10], а именно:

– организационную невозможность переноса данных (полного или частичного) за периметр организации (влечет невозможность использования ETL) для некоторых ИС;

– обеспечение высокой скорости доступа к консолидированным данным (ограничение на использование EII);

– обеспечение доступа пользователей к расчетным подсистемам, функционирующим в рамках некоторых ИС (диктует необходимость консолидации пользовательских интерфейсов (необходимость использования EAI)).

Решить данные проблемы можно путем создания методологии интеграции ИС, сочетающей использование трех обозначенных выше методов интеграции (рис. 7) для обеспечения наиболее тесной консолидации разнородных ИС по свойствам неорганических веществ.

Проиллюстрируем на примерах созданной интегрированной ИС по свойствам неорганических веществ и материалов подходы к решению обозначенных выше проблем.

Рис. 7. Методология интеграции ИС

Fig. 7. The IS integration methodology

EII для интеграции данных по свойствам неорганических веществ

Метод создания хранилищ данных хорошо зарекомендовал себя для решения проблем интеграции в рамках одной организации. Поскольку интегрируемые ИС по свойствам неорганических веществ создавались в рамках различных организаций (и даже стран), существовали организационные проблемы, связанные с запретом копирования данных для их перемещения в централизованное хранилище. В то же время разрешался программный (и, как правило, ограниченный) доступ к данным через внешний API. Поэтому для интеграции в этих случаях использовался метод EII совместно с подходом Local-as-View, который рассматривает схемы локальных источников данных как материализованные представления в терминах общей глобальной схемы предметной области. Сама глобальная схема описана в терминах иерархии химических понятий «система–вещество–модификация» [10], а программные адаптеры, размещенные поверх интегрируемых источников данных, преобразуют их внутренние информационные структуры к глобальной схеме.

Для осуществления кроссплатформенного взаимодействия используется технология web-сервисов, то есть все программные адаптеры реализуются в качестве web-сервисов, доступных по протоколу SOAP, что позволяет успешно разрешать платформенные конфликты, так как 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, и имеют одинаковое, стандартизированное в рамках общей схемы WSDL-описание, что обеспечивает унифицированную работу со всеми подобными web-адаптерами со стороны предметного посредника.

С учетом общей схемы предметной области разработано унифицированное описание web-сервисов программных адаптеров на языке WSDL, которому должны удовлетворять все web-сервисы адаптеров интегрируемых ИС: http://crystal.imet-db.ru/eii_crystal/eii_crystal.asmx?WSDL. В итоге, например, адаптер поверх ИС «Кристалл» (crystal.imet-db.ru) http://crystal.imet-db.ru/eii_crystal/eii_crystal.asmx работает по тем же принципам и предоставляет тот же API, что и адаптер для ИС «BandGap» (bg.imet-db.ru) http://bg.imet-db.ru/eii_bandgap/eii_bandgap.asmx. В этом смысле система является хорошо масштабируемой и расширяемой – достаточно написать адаптер, удовлетворяющий указанному выше WSDL-описанию, и зарегистрировать его адрес в каталоге EII-интеграции, с которым работает предметный посредник. Сам посредник тоже реализован в виде web-сервиса, доступного по адресу http://meta.imet-db.ru/eii/service.asmx.

ETL для создания хранилища данных в рамках организации

В случае необходимости обеспечения быстрого и надежного доступа к интегрированному источнику данных оптимальным является создание хранилища данных, в которое помещается требуемая информация. Создание подобного хранилища, как правило, возможно только в рамках одной организации. В качестве примера приведем хранилище данных, созданное в рамках ИМЕТ РАН для консолидации данных по свойствам химических элементов, содержащихся ранее в БД «Элементы» и БД «Фазы» по основным свойствам неорганических соединений. В итоге по ETL-методу получено единое хранилище, функционирующее под управлением Microsoft SQL Server, используемое информационно-аналитической системой для компьютерного конструирования неорганических соединений. Более того, несколько выходя за рамки классического ETL-метода, были консолидированы и соответствующие web-приложения в рамках единого домена: ИС «Фазы» доступна по адресу http://phases.imet-db.ru/, а ИС «Элементы» по адресу http://phases.imet-db.ru/elements/.

EAI для интеграции пользовательских интерфейсов

При интеграции web-приложений ИС по свойствам веществ необходим поиск релевантной информации – http://meta.imet-db.ru/service/service.asmx.

Интеграция пользовательских интерфейсов интегрируемых ИС по свойствам неорганических веществ позволяет использовать расчетные подсистемы, функционирующие только в контексте исходной ИС. Для обеспечения возможности перехода пользователя в контекст другой ИС для просмотра информации был

разработан web-сервис (http://meta.imet-db.ru/service/service.asmx), являющийся средством поиска релевантной информации в контексте интегрированной системы. Поиск осуществляется в метабазе, содержащей сведения о химических сущностях, описанных в интегрируемых ИС. Актуальность сведений в метабазе поддерживается за счет использования web-сервиса обновления: http://meta.imet-db.ru/muservice/muservice.asmx. Таким образом, пользователь, находясь в контексте одной из ИС, может перейти в другую для просмотра релевантной информации, например, при просмотре сведений по ниобату лития в ИС «Кристалл» предоставляется возможность перехода в ИС «BandGap», «AtomWork», «ТКВ» непосредственно на страницу с $LiNbO_3$.

Очень часто при поиске данных по свойству того или иного вещества неискушенный пользователь не знает, к какой ИС по свойствам неорганических веществ стоит прибегнуть для получения информации. Поэтому была создана специализированная ИС, позволяющая потребителю данных по свойствам неорганических веществ получить возможность просмотра связанной информации по свойствам заданной химической системы в разных ИС: http://meta.imet-db.ru/. Данная ИС является web-приложением ASP.Net, написанным на языке C# (клиентская часть: DHTML, CSS, AJAX, jQuery) с использованием ADO.Net, для доступа к метабазе, содержащей консолидированные метаданные. Для построения запросов используются языковые средства Transact-SQL, являющегося диалектом языка SQL.

В настоящее время интеграция на уровне web-приложений проведена для ИС «Bandgap», «Кристалл» (русско- и англоязычные версии), «Диаграмма», «Фазы», «Элементы», «Кремний» (все ИМЕТ РАН), «Термические константы веществ» (ТКВ, ОИВТ РАН совместно с МГУ) и «AtomWork» (бывшая Pauling File, разработанная в NIMS, Япония).

Заключение

Полная актуальность данных может быть достигнута только за счет использования методов EII либо EAI. Кроме того, преобразование данных этими методами осуществляется в рамках ИС источников данных. Таким образом, смена форматов исходных данных отражается на интегрированной ИС в минимальной степени.

Использование метода хранилищ данных (ETL) предлагается для создания интегрированного источника данных в рамках одной организации. Это позволит достичь максимальной надежности и скорости работы с интегрированными данными со стороны систем компьютерного конструирования неорганических соединений или других высокоуровневых средств интеграции. Использование метода интеграции данных (EII) предлагается для виртуальной интеграции материаловедческой информации между ИС, как правило, относящимися к разным организациям, запрещающим физическое копирование данных или предоставляющим ограниченный доступ к данным на платной основе. Таким образом, на нижнем уровне (в рамках организации) данные интегрируются с помощью хранилищ данных (ETL), а затем на более высоком уровне интеграция осуществляется с использованием метода EII (см. http://www.swsys.ru/uploaded/image/2020-2/2020-2-dop/3.jpg). Заметим, что возможна реализация многоуровневой схемы использования хранилищ данных и виртуальной интеграции для обеспечения требуемой скорости обработки и масштабируемости.

Литература

1. Imhoff C. Understanding the Three E's of Integration EAI, EII and ETL. DM Review Magazine, 2005, iss. 4. URL: http://www.dmreview.com/article_sub.cfm?articleId=1023893 (дата обращения: 25.01.2020).

2. Масютин В.В., Дударев В.А. Системный анализ технологий интеграции гетерогенных баз данных // Новейшие достижения европейской науки: матер. VII Междунар. науч.-практич. конф. 2011. Т. 34. С. 35–36.

3. Sherman R. Business intelligence guidebook: From data integration to analytics. Morgan Kaufmann Publ., 2014, 550 p.

4. Inmon W.H. Building the data warehouse. John Wiley Publ., 2005, 576 p.

5. Bouaziza S., Nablib A., Gargouric F. Design a data warehouse schema from document-oriented database. Procedia Computer Science, 2019, vol. 159, pp. 221–230.

6. Themistocleous M. Justifying the decisions for EAI implementations: a validated proposition of influential factors. J. Enterprise Information Management, 2004, vol. 17, no. 2, pp. 85–104.

7. Gericke A., Klesse M., Winter R., Wortmann F. Success factors of application integration: an exploratory analysis. Communications of the Association for Information System, 2010, vol. 27, no. 1, pp. 677–694. DOI: 10.17705/1CAIS.02737.

8. Волкова В.Н., Денисов А.А. Теория систем и системный анализ. М., 2012. 679 с.

9. Huang C., Cai H., Xu L., Xu B., Gu Y., Jiang L. Data-driven ontology generation and evolution towards intelligent service in manufacturing systems. Future Generation Computer Systems, 2019, vol. 101, pp. 197–207. DOI: 10.1016/j.future.2019.05.075.

10. Дударев В.А. Интеграция информационных систем в области неорганической химии и материаловедения. М.: КРАСАНД, 2016. 320 с.

Software & Systems Received 30.01.20
DOI: 10.15827/0236-235X.130.283-296 2020, vol. 33, no. 2, pp. 283–296

Integration methods analysis for the development of information-analytical systems on inorganic substances properties

V.A. Dudarev [1], *Ph.D. (Engineering), Associate Professor, vdudarev@hse.ru*
I.O. Temkin [2], *Dr.Sc. (Engineering), Professor, igortemkin@yandex.ru*
V.F. Kornyushko [3], *Dr.Sc. (Engineering), Professor, vfk256@mail.ru*

[1] *National Research University Higher School of Economics, Moscow, 109028, Russian Federation*
[2] *National University of Science and Technology «MISIS», Moscow, 119049, Russian Federation*
[3] *Moscow Technological University (MIREA), Moscow, 119454, Russian Federation*

Abstract. The paper considers the system analysis application to formalize information processes that occur in integrated information systems (IS) and develop a system of recommendations for choosing one of the integration methods for solving practical IS consolidation problems.

The analysis starts with the identification of the basic information processes occurring in local IS and continues with the studying of their transformation in integrated systems built using three consolidation methods: Enterprise Application Integration (EAI), Enterprise Information Integration (EII) and software for extracting, converting, and loading data (Extract, Transform, Load – ETL), based on data warehouse technology.

The authors synthesized a generalized scheme of integration methods for heterogeneous information systems, and based on it, compared the methods using ten criteria. Based on the comparison results, the authors offered recommendations for choosing the preferred method of heterogeneous information systems integration.

There is a possibility to use ETL and EII methods together to achieve the optimal performance of the integrated data source. Since none of the methods covers all the needs arising from the consolidation of information systems, there is a methodology for their joint use for the closest integration of information systems.

Briefly, the paper reviews the practical application results of the developed integration methodology for the IS consolidation in the inorganic materials science field, both domestic and foreign: there is a created web-service set, which is the main technical means that provides interaction of distributed cross-platform information systems on the properties of inorganic substances for data consolidation, and the user interface of a single entry point www.imet-db.ru, allows any user to search and view information on the properties of substances contained in integrated information systems.

Keywords: nformation processing, data integration, application integration, integration methods comparison, criteria for integration method selection.

References

1. Imhoff C. *Understanding the Three E's of Integration EAI, EII and ETL*. DM Review Magazine, 2005, iss. 4. Available at: http://www.dmreview.com/article_sub.cfm?articleId=1023893 (accessed January 25, 2020).

2. Masyutin V.V., Dudarev V.A. System analysis of heterogeneous database integration technologies. *Proc. VII Intern. Sc. Conf*., Sofia, 2011, vol. 34, pp. 35–36 (in Russ.).

3. Sherman R. *Business Intelligence Guidebook: From Data Integration to Analytics*. Morgan Kaufmann Publ., 2014, 550 p.

4. Inmon W.H. *Building the Data Warehouse*. *Fourth Edition*. John Wiley Publ., 2005, 576 p.

5. Bouaziza S., Nablib A., Gargouric F. Design a data warehouse schema from document-oriented database. *Procedia Computer Science*, 2019, vol. 159, pp. 221–230.

6. Themistocleous M. Justifying the decisions for EAI implementations: a validated proposition of influential factors. *J. Enterprise Information Management*, 2004, vol. 17, no. 2, pp. 85–104.

7. Gericke A., Klesse M., Winter R., Wortmann F. Success factors of application integration: an exploratory analysis. *Communications of the Association for Information System*, 2010, vol. 27, no. 1, pp. 677–694. DOI: 10.17705/1CAIS.02737.

8. Volkova V.N., Denisov A.A. *Systems Theory and System Analysis.* Moscow, 2012, 679 p. (in Russ.).

9. Huang C., Cai H., Xu L., Xu B., Gu Y., Jiang L. Data-driven ontology generation and evolution towards intelligent service in manufacturing systems. *Future Generation Computer Systems*, 2019, vol. 101, pp. 197–207. DOI: 10.1016/j.future.2019.05.075.

10. Dudarev V.A. *Integration of Information Systems in the Field of Inorganic Chemistry and Materials Science*. Moscow, 2016, 320 p. (in Russ.).

Для цитирования

Дударев В.А., Темкин И.О., Корнюшко В.Ф. Анализ методов интеграции для разработки информационно-аналитических систем по свойствам неорганических соединений // Программные продукты и системы. 2020. Т. 33. № 2. С. 283–296. DOI: 10.15827/0236-235X.130.283-296.

For citation

Dudarev V.A., Temkin I.O., Kornyushko V.F. Integration methods analysis for the development of information-analytical systems on inorganic substances properties. *Software & Systems*, 2020, vol. 33, no. 2, pp. 283–296 (in Russ.). DOI: 10.15827/0236-235X.130.283-296.