

В. А. Дударев, канд. техн. наук, доцент Московского государственного университета тонких химических технологий им. М. В. Ломоносова, vic@imet.ac.ru

В. В. Масютин, аспирант Московского государственного университета тонких химических технологий им. М. В. Ломоносова, inftech@mitht.ru

Интеграция автономных источников данных для исследования свойств полупроводниковых материалов

Интеграция информационных ресурсов по свойствам полупроводниковых материалов, разработанных в различных странах, необходима для создания новых материалов с требуемыми свойствами. Для интеграции автономных источников данных предлагается использование федеративного подхода, при котором обеспечивается интеграция схем хранения данных и доступ к физически распределенным данным. Представлены модели потоков данных и функциональные модели интеграции автономных источников данных.

Ключевые слова: интеграция автономных источников данных, информационное моделирование.

Введение

В настоящее время во всех промышленно развитых странах ведется разработка специализированных информационных систем (ИС) по свойствам полупроводниковых материалов. Наиболее мощные информационные системы, основанные на современных программно-аппаратных платформах, предлагают NIST (National Institute of Standards and Technology — Национальный институт стандартов и технологий, США) [1] и NIMS (National Institute of Materials Science — Национальный институт материаловедения, Япония) [2]. В России лидером в области создания информационных ресурсов по свойствам полупроводниковых материалов является Институт металлургии и материаловедения им. А. А. Байкова РАН (ИМЕТ РАН) [3].

Ни одна из существующих ИС по свойствам неорганических веществ и материалов не может предоставить исчерпывающую информацию обо всей совокупности

свойств конкретного вещества. Фрагментарность данных о полупроводниковых материалах — одна из проблем, усложняющая создание новых материалов с требуемыми свойствами [4]. Один из путей решения проблемы — интеграция информационных ресурсов. Актуальность интеграции была осознана в последние годы не только на национальном, но и на международном уровне, что привело к созданию специальной международной комиссии (Materials Task Group) [5], занимающейся выработкой стандартов для интеграции материаловедческих ИС. Однако, несмотря на предпринимаемые усилия, говорить об успехах в этой области преждевременно.

Методы интеграции данных

Под интеграцией данных обычно понимают комбинирование данных из разных источников с тем, чтобы создать их целостное, унифицированное, согласованное

представление [6]. Отметим, что информация в различных информационных системах может храниться не только в форме распределенных баз данных, но и в других видах, например в форматах электронных таблиц (например, Microsoft Excel), текстовых, бинарных проприетарных, формате XML и т. п. Все это значительно затрудняет интеграцию разнородных ИС.

Наибольшее распространение получили два базовых метода интеграции данных: консолидация и федерализация.

При использовании метода консолидации данные собираются из нескольких первичных систем и интегрируются в одно постоянное хранилище. Такое хранилище может быть использовано как источник данных для приложений, выполняющих формирование поисковых запросов.

Федеративный подход обеспечивает единое виртуальное представление разнородных источников данных. При этом данные хранятся в разных по составу и структуре источниках, информация в которых может частично дублироваться. Источники остаются полностью автономными. Интеграция данных сводится к интеграции схем хранения и созданию программного компонента (посредника), обеспечивающего прозрачный доступ к физически распределенным данным. В отношении конечного приложения взаимодействие осуществляется с единой базой данных в едином стандартизированном формате. Интеграция корпоративной информации (Enterprise information integration, EII) — пример технологии, которая поддерживает федеративный подход к интеграции данных.

Основным преимуществом федеративного подхода является то, что открыт доступ к текущим данным, без задержек в обновлении, в отличие от подхода, основанного на создании единого хранилища данных. Этот подход избавляет также от необходимости копировать данные из источника в хранилище. Поэтому федерализация данных применяется в тех случаях, когда политика безопасности данных и лицензионные

ограничения запрещают копирование данных первичных систем.

Среди недостатков подхода — нелинейно возрастающая сложность реализации при увеличении числа источников (наличие существенных различий в модели данных может значительно усугубить ситуацию), высокие требования к качеству связи. При выполнении запросов могут возникать задержки, связанные с необходимостью обмена данными между источниками и программ-посредником [7].

Неформальное описание интеграции автономных источников данных

Выполнение внешнего запроса при использовании автономных источников данных осуществляется следующим образом. Программа-посредник анализирует поступающие от приложений запросы, формирует производные запросы к различным источникам, агрегирует полученные данные и возвращает ответ приложению. Для взаимодействия между посредником и приложениями используется единый, стандартизированный в рамках данной системы интеграции данных, интерфейс. Важный элемент федеративной ИС — метаданные, представляющие собой описания источников данных. Метаданные используются посредником для выбора источников данных и формирования запросов к ним. Для каждого из источников дополнительно реализуется адаптер извлечения и преобразования данных, который обеспечивает учет особенностей первичной ИС. На этапе агрегации возможно преобразование и изменение данных, устранение конфликтов данных.

Применение методов системного анализа, таких как декомпозиция и иерархическое упорядочение, позволяет построить неформальное описание интеграции автономных источников данных. Это описание приведено на рис. 1 и будет в дальнейшем детализировано и формализовано с различных точек зрения.

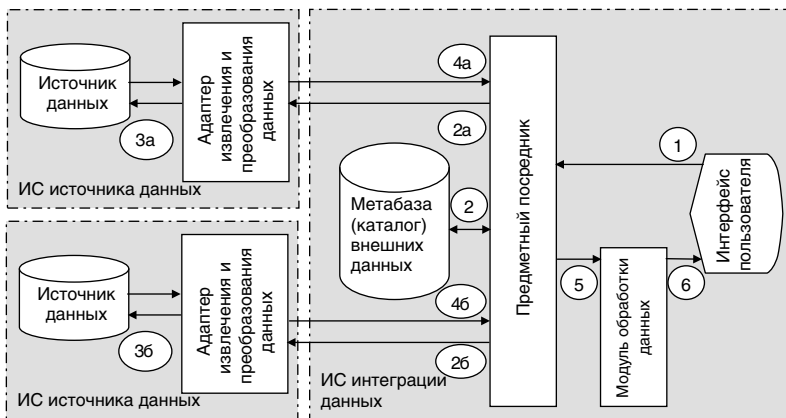


Рис. 1. Интеграция автономных источников данных — неформальное описание

Информационные процессы: 1 – формирование запроса к данным; 2 – определение данных для извлечения; 3 – извлечение данных и преобразование к одному формату; 4 – интеграция данных, полученных из разных источников; 5 – обработка данных; 6 – представление данных.

Комплекс информационных моделей для ИС интеграции данных

Моделирование федеративного подхода к интеграции данных включает в себя следующие этапы [8]:

- моделирование потоков данных в ИС интеграции данных;
- функциональное моделирование процессов обработки информации при интеграции данных;
- моделирование данных для обмена информацией с ИС источников данных.

В статье рассматриваются первые два этапа.

Для формализованного описания потоков данных в ИС интеграции данных используются диаграммы потоков данных (Data Flow Diagram, DFD) [9]. Моделирование потоков данных начинается с построения обобщенной диаграммы, представленной на рис. 2. Единственный информационный процесс на этой диаграмме реализует основное назначение ИС — интеграция и обработка данных. Абстрактный накопитель «источники данных» представляет собой информационные системы, в которых хранится первичная информация и которые подлежат интеграции при помощи одной из технологий интеграции данных. Абстрактный

накопитель «данные для пользователя» — результат работы ИС интеграции данных. Возможное дальнейшее применение полученных данных лежит за рамками ИС интеграции данных и на контекстной диаграмме не отражено. Абстрактный накопитель «метаданные» содержит описания источников данных.

Далее эта диаграмма детализируется. Выделяют три последовательных информационных процесса: извлечение, преобразование и представление данных. Для взаимодействия этих процессов используются промежуточные накопители данных. Диаграмма, полученная на 1-м уровне декомпозиции, представлена на рис. 3.

Для детализации описания процессов используется методология функционального моделирования IDEF0 [10]. Обобщенная функциональная модель очень близ-

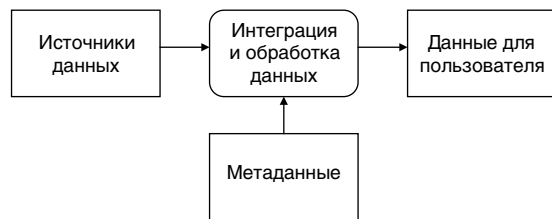


Рис. 2. Обобщенная диаграмма потоков данных в ИС интеграции данных

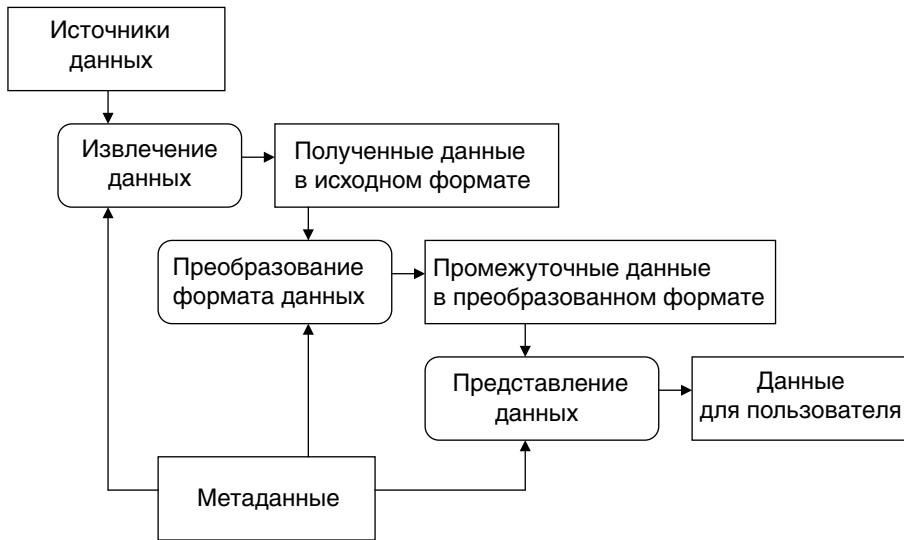


Рис. 3. Диаграмма потоков данных первого уровня декомпозиции

ка к обобщенной модели потоков данных, но при дальнейшей детализации используются выразительные возможности функционального моделирования, в частности возможность указания роли при описании взаимосвязей процессов. На функциональной диаграмме процессы изображаются в виде функциональных блоков, а взаимосвязи между ними — с помощью стрелок, причем

положение стрелки относительно функционального блока определяет его роль.

Функциональная диаграмма 1-го уровня декомпозиции представлена на рис. 4. Она состоит из четырех функциональных блоков, связанных отношениями «вход-выход»: результат одного процесса является исходными данными для другого. Последовательностью блоков задается последовательность

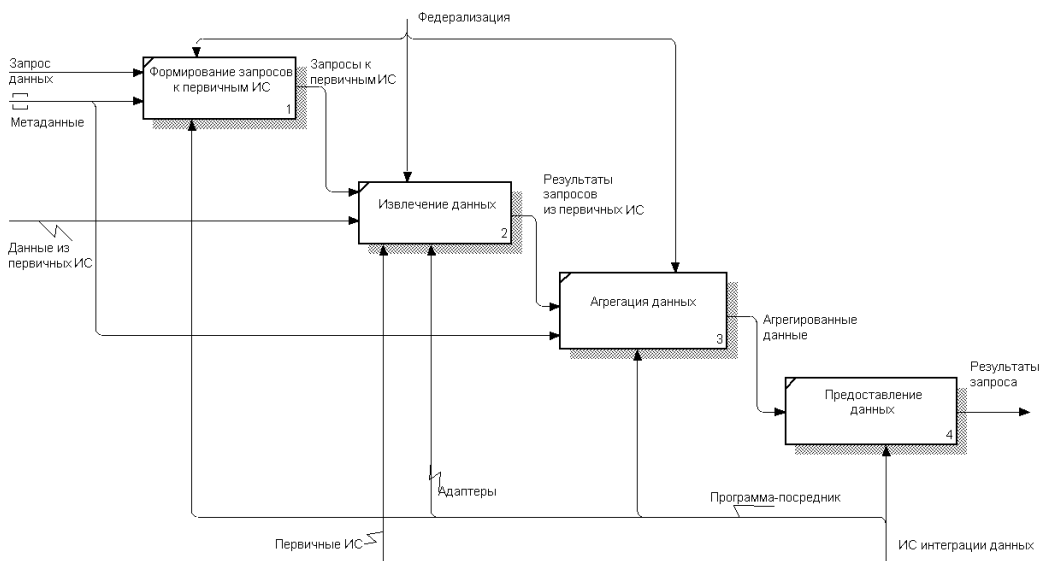


Рис. 4. Функциональная диаграмма первого уровня декомпозиции

выполнения информационных процессов: формирование запросов к первичным ИС, извлечение данных, агрегация данных, предоставление данных пользователю. Стрелки «запрос данных» и «данные из первичных источников» описывают исходные данные для всего процесса интеграции. Стрелка «метаданные» является туннельной — она не представлена на родительской диаграмме, так как состав метаданных становится понятен в процессе детализации. Стрелки с ролью «механизм» (ведущие к блокам от нижнего края диаграммы) характеризуют исполнителей процессов. Так, первичные ИС используются на этапе извлечения данных. Для описания составных частей ИС интеграции данных используется ветвление стрелки. Так, стрелки, задающие программу — посредник и адаптеры, являются частью стрелки «ИС интеграции данных».

Заключение

Применение федеративного подхода позволяет интегрировать как российские, так и зарубежные информационные ресурсы по свойствам полупроводниковых материалов без физического перемещения данных в хранилище. Задачу взаимодействия с источниками данных реализует программа-посредник, используя метаданные об организации данных в информационных ресурсах. Взаимодействие посредника с источни-

ками данных описано в виде модели потоков данных и функциональной модели.

Список литературы

1. Интернет-источник <http://www.nist.gov>.
2. Интернет-источник <http://www.nims.go.jp/eng>.
3. Интернет-источник <http://www.imet.ac.ru>.
4. Киселева Н. Н. Компьютерное конструирование неорганических соединений: использование баз данных и методов искусственного интеллекта // М.: Наука, 2005. — 289 с.
5. Интернет-источник <http://www.codata.org/taskgroups/TGmatsdata/index.html>.
6. Черняк Л. Интеграция данных: синтаксис и семантика // Открытые системы. 2009. № 10.
7. Граничин О. Н., Кияев В. И. Информационные технологии в управлении. СПб.: Бином, 2008. — 336 с.
8. Захаров В. Н., Калиниченко Л. А., Соколов И. А., Ступников С. А. Конструирование канонических информационных моделей для интегрированных информационных систем // Информатика и ее применения. 2007. Т. 1. Вып. 2.
9. Бурляева Е. В., Колыбанов К. Ю., Панова С. А. Информационная поддержка систем принятия решений на производственных предприятиях химического профиля. М.: Издательство МИТХТ им. М. В. Ломоносова, 2013. — 196 с.
10. Рекомендации по стандартизации. Информационные технологии поддержки жизненного цикла продукции. Методология функционального моделирования — Госстандарт России. М.: Издательство стандартов, 2001.

V. Dudarev, PhD in Technique, Associate Professor, Lomonosov Moscow State University of Fine Chemical Technologies, vic@imet.ac.ru

V. Masjutin, Graduate Student, Lomonosov Moscow State University of Fine Chemical Technologies, inftech@mitht.ru

Autonomous data sources integration for the semiconductor materials properties study

Integration of information resources on the properties of semiconductor materials developed in different countries is necessary for creation of new materials with the required properties. Federated approach to autonomous data sources integration is suggested. The approach allows the integration of data storage and access of physically distributed data. The data flow model and functional models of autonomous data sources integration are developed.

Keywords: autonomous data sources integration, information modeling.