РАЗВИТИЕ БАНКА ДАННЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ ВИНИТИ ПО ХИМИИ И ХИМИЧЕСКОЙ ТЕХНОЛОГИИ: ОТ СТРУКТУРНЫХ ДАННЫХ К МАТЕРИАЛЬНОЙ ХИМИИ И ТЕХНОЛОГИИ. 2. ПРОГРАММНО-ТЕХНОЛОГИЧЕСКИЕ РЕШЕНИЯ РЕГИСТРОВ ХИМИЧЕСКИХ СОЕДИНЕНИЙ И РЕАКЦИЙ*

Н.Н. Афонина, А.Л. Бирюков, Д.И. Гончаренко, В.М. Зацепин, В.А. Иванченко, Н.Н. Кочанова, А.К. Мартиросов, О.М. Нефедов, Р.В. Остапчук, О.С. Сафронова, М.А. Федоровская

Всероссийский институт научной и технической информации РАН, Москва
Научно-инженерный центр «СИСТЕХ», Москва (systech@aha.ru)

Рассмотрены вопросы развития банка данных информационных ресурсов ВИНИТИ по химии и химической технологии для адекватной современной поддержки фундаментальных и инновационно-технологических разработок в областях химии и химической технологии. Разработаны и подготовлены к внедрению информационные и программно-технологические решения по регистру химических веществ и реакций, АРМ для содержательной переработки первоисточников, подготовки рефератов и файлов данных по структурам и реакциям для загрузки в базы данных хранилища данных, а также АРМ редактора-индексатора для систематизации номенклатурных названий химических соединений и восстановления пропущенных данных с использованием конверторов структура–название CAS/IUPAC. Созданы прототипы баз данных регистров химических соединений, реакций и смесей, химико-технологических схем, процессов, аппаратов и аналитических методов. Разработаны и подготовлены к внедрению новые информационные и программно-технологические решения для формализованного, в том числе и графического, представления сведений о химическом производстве.

Ключевые слова: химические информационные объекты, химические структуры, химические реакции, химические композиции, химические продукты, химико-технологические схемы, процессы и аппараты, системы баз данных, WWW-доступ.

The issues of development of VINITI information resources databank on chemistry and chemical technology for adequate modern support of fundamental and innovation technique development in the field of chemistry and chemical technology are considered. The complex of new information and software-technological means for Register of Chemical Substances and Reactions, CWP for pithy processing of primary sources, preparing of abstracts and data files on structures (SDF) and reactions for loading in databases, as well as the editor-indexer CWP for systematization of nomenclature names of chemical compounds and regeneration of missed data using of structure–CAS/IUPAC name converters, have been developed and prepared to practical using. The prototypes of databases of registers of chemical compounds, reactions and mixtures, chemical-engineering schemes, processes, apparatuses and analytical methods are created. The new information and software-technological means of formalized, including graphic, presenting data on chemical production are developed and prepared to practical using.

Keywords: chemical information objects, chemical structures, chemical reactions, chemical compositions, chemical products, chemical-technological process flow diagrams, processes and apparatuses, databases systems, WWW-access.

Программно-технологические решения проблемы формализации и регистрации химико-структурных данных и создания регистров химических соединений и реакций обоснованы нами ранее как важнейший узловой этап развития банка данных информационных ресурсов ВИНИТИ по химии и химической технологии [1–3]. При этом в рамках разработанной кон-

* Работа частично поддержана грантом РФФИ 01-07-90097в.

© ВИНИТИ, 2002

цепции моделей и баз данных (БД) традиционная составляющая информационного ресурса ВИНИТИ по химии и химической технологии – база структурных данных (СД) может рассматриваться только как промежуточный ресурс, требующий дополнительной реорганизации, с целью обеспечения современного уровня представления информации в области химии и химической технологии, характеризующейся обширной многомерной фактографией (более 20 миллионов химических структур со своими уникальными названиями, различными синонимами названий и регистрационными кодами, химическими реакциями, физико-химическими, медико-биологическими, токсикологическими и другими свойствами). В должной степени этим требованиям отвечает разрабатываемая нами адаптивная (развивающаяся) система информационного обеспечения (АСИО) работ в соответствующей области знаний [2, 3].

В данной работе представлены результаты исследований по второму этапу формирования АСИО [1], в рамках которого разрабатывались информационные и программно-технологические средства переработки входного потока информации, ориентированные на реорганизацию БД ВИНИТИ с использованием регистров химических соединений и реакций, и создания современного корпоративного информационно-аналитического ресурса (хранилища данных).

Методологическую основу разработки современных БД – хранилищ данных, ориентированных на проведение серьезных информационно-аналитических изысканий, включая поддержку принятия решений по приоритетным направлениям науки и техники [4–8], составляют классификация, стандартизация, кодификация и регистрация информационных объектов.

В области химии наиболее представительной и актуальной системой регистрации химико-структурных данных является CAS – Chemical Abstracts Service [9], БД которого включает более 21 миллиона веществ и 25 миллионов записей с ежедневным пополнением порядка 4 тысяч новых соединений. В основе регистрации лежит присвоение каждому новому химическому соединению уникального цифрового номера (до 9 знаков, разделенных на три части дефисами), в определении которого, по заявлению CAS, хотя и принимают участие ученые CAS, но для внешних пользователей представляющего не более, чем порядковый номер химиката в БД CAS. Сложившаяся за последние годы монополия CAS на оказание платной услуги на присвоение порядкового номера, не несущего никакой химической информации, во многих случаях затрудняет его использование в качестве кода/указателя для обращения к информации.

В качестве альтернативы номеру CAS в научных коммуникациях все более значимую роль играет линейный код, базирующийся на переводе химической структуры в уникальную строку символов, допускающую унифицированную генерацию регистрационного хэш-кода. Здесь выделяются такие популярные линейные коды, как SMILES [10] и SYBYL [11]. В плане программной реализации регистрации химических структур в БД можно выделить подходы, предлагаемые фирмами MDL [12] и DayLight [13] с более открытой и масштабируемой методологией. На первом этапе регистрации новому веществу по стандартному набору хэш-кодов сопоставляются вещества из БД, а на втором с помощью анализа молекулярных дескрипторов (метода «отпечатка пальца») подтверждается уникальность регистрируемого вещества. Различие подходов MDL и DayLight состоит в использовании для «отпечатка пальца» определенного набора дескрипторов и набора постоянно генерируемых дескрипторов, соответственно.

Поэтому на этапе отработки программно-технологических решений по созданию регистров химических соединений и реакций мы исходили из принципов формирования информационного ресурса – хранилища химико-структурных и фактографических данных по результатам содержательной переработки входного потока информации ВИНИТИ. При этом фактография, в том числе молекулярные дескрипторы, так же как и непосредственно молекулярные представления химической структуры и реакции, являются способом отражения – измерения реальных химических объектов в сфере материальной химии.

Отметим, что достигнутый в последние годы прогресс в компьютерных технологиях, связанных с химической информатикой, часто выражается и в неоправданном смешении понятий о химическом веществе (реальном объекте материальной химии) и химической структуре

 © ВИНИТИ, 2002

(идеальном объекте – факте, результате его отражения, измерения) [1].

Для отработки информационных и программно-технологических средств регистров химических соединений и реакций использовали реляционную СУБД MS SQL. В качестве инструмента для разработки универсального пользовательского интерфейса БД (системы АРМ – клиентских приложений пользователя) использовали интегрированную среду разработки программных приложений Delphi/C++.

1. Редактор структурных данных

Принципы построения крупного хранилища химико-структурных данных как в отношении данных и форматов их хранения, так и в отношении запросов, включая возможности манипулирования обобщенными (Маркуш-формулами) представлениями химических структур, исключают возможность использования в качестве базового программного средства любого из множества свободно распространяемых и отдельно коммерчески доступных редакторов молекулярных структур [12–15]. В должной мере этим принципам удовлетворяет разработанный в НИЦ «СИСТЕХ» редактор структурных данных MSE, важными характеристиками которого являются:

- открытые для генератора информационного ресурса программные коды и форматы данных;
- поддержка стандартных форматов (MDL и др.) представления молекулярных данных, в том числе формул Маркуша, дополненных SMILES/SLN линейными кодами определения переменных заместителей;
- пополняемые библиотеки молекулярных фрагментов – типовых структур;
- пополняемые библиотеки поименованных (IUPAC/синонимы) Маркуш-заместителей;
- редактирование структурной части, названия и логических условий определения переменных заместителей, в том числе с использованием набора встроенных функций;
- развитые средства ввода/редактирования, поиска, классификации и регистрации химико-структурных объектов в БД с использованием формул Маркуша;
- Windows 2000/NT.

Редактор ориентирован как на работу с любыми структурными, в том числе патентными БД, так и на задачи разметки/индексации химических баз данных для классификации и синтеза информации.

На рис. 1 приведена экранная форма редактора MSE на этапе ввода - формализации обобщенных химических структур с использованием формул Маркуша. Дополнительное окно включает список определений переменных заместителей – радикалов, используемых в текущей структуре.

На рис. 2 приведена экранная форма редактора MSE на этапе ввода-формализации химической реакции как наборов молекул исходных реагентов и конечных продуктов.

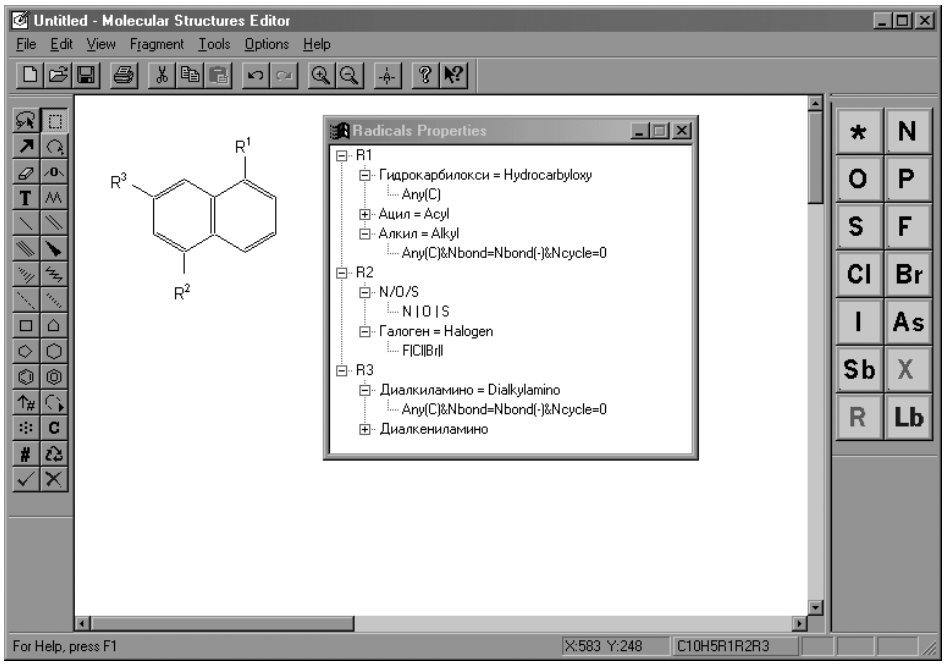

Рис. 1. *Редактор структурных данных MSE: формализация обобщенных химических структур с использованием формул Маркуша*

2. Автоматизированные рабочие места (АРМ) для содержательной переработки первоисточников и формализации химических соединений и реакций

В соответствии с развиваемой концепцией АСИО операционно-технологическая база данных (ОТБД) и автоматизированные рабочие места (АРМ) – клиентские приложения реализуют процедуры экстракции, формализации, унификации/стандартизации химико-структурных и других фактографических данных из обрабатываемых документов-первоисточников. Особенностями программно-технологических решений по АРМ для работы с соответствующими оперативными данными являются:

- ориентация на создание мощной документально-фактографической химико-структурной БД хранилища данных информационного ресурса;
- унифицированная формализация экстрагируемых из первоисточников химико-структурных данных с использованием стандартных библиотек молекулярных структур и их фрагментов, а также динамически формируемых локальных выборок – справочников структур, в том числе из химического регистра хранилища данных, и включения их в состав локальных АРМ;
- возможность формализованного представления, загрузки и работы с документальными и реферативно-библиографическими данными различных типов обрабатываемых документов с гибко настраиваемой структурой элементов данных, в том числе и в соответствии с действующими научно-техническими предписаниями (НТП) ВИНИТИ;
- возможность экспорта оперативных данных в стандартных файловых форматах (ISO, SDF, RDF и др.);
- возможность интеграции в технологический цикл содержательной переработки информации других известных программных и информационных продуктов: MDL (ISIS-Base) [12], ACD/Labs (Index Name) [14], CambridgeSoft (ChemFinder) [15] и др.

На рис. 3 приведена экранная форма АРМ для работы с реферативно-библиографическими данными документа, обрабатываемого в ОТБД. Каждому типу обрабатываемого документа может соответствовать несколько форм сочетания/отображения его элементов данных (основная – по НТП, промежуточные – по технологическим стадиям и оригинал-макетам создаваемых информационных продуктов). Отображаемые на экранной форме химические структуры и реакции представляют специальные элементы данных обрабатываемого документа – химические информационные объекты, работа с которыми осуществляется в соответствующих экранных формах АРМ (рис. 4 и 5).

На рис. 4 приведена экранная форма АРМ для работы с данными по химическим структурам, включающая возможности:

- просмотр/редактирование химической структуры (левое верхнее окно) для текущего соединения

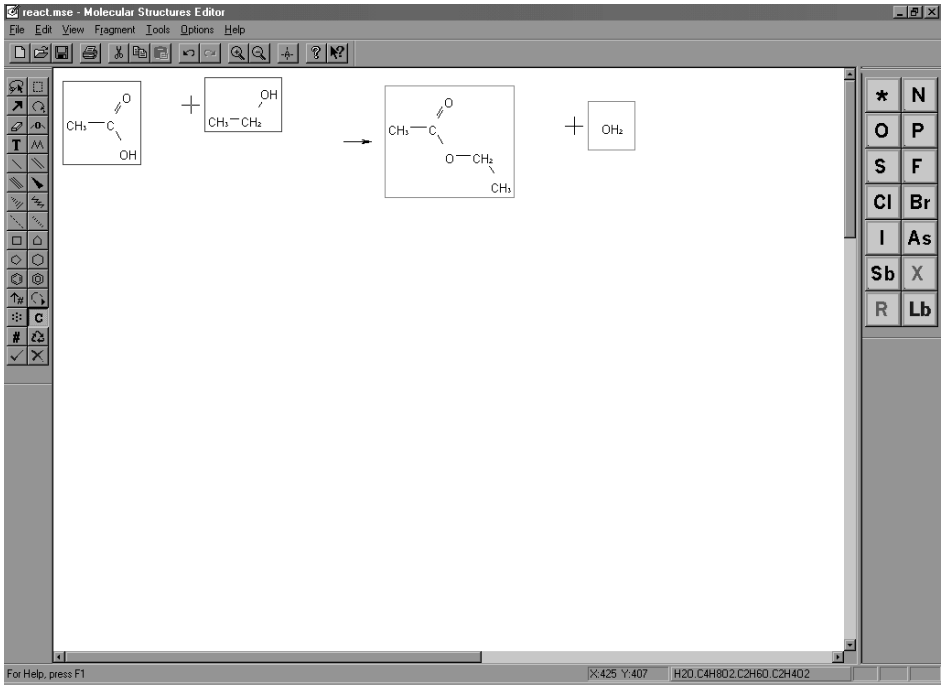

Рис. 2. *Редактор структурных данных MSE: ввод химической реакции*

 © ВИНИТИ, 2002

из списка уже формализованных в текущем документе химических соединений (правое нижнее окно) или из общего списка справочника соединений локального АРМ (правое верхнее окно);

• выбор из списка справочника соединения с «подходящей структурой» и привязка к текущему обрабатываемому документу (добавление к списку формализованных соединений), а при отсутствии такового – внесение в общий список локального справочника новой записи и ее связывание с документом;

• формирование списка унифицированных записей термов свойств (левое нижнее окно) текущего соединения из списка уже формализованных в обрабатываемом документе химических соединений.

Важно подчеркнуть, что ведение в рамках АРМ справочника соединений позволяет в определенной мере унифицировать ввод данных и исключить дублирование химических структур. При этом справочники актуализируются как за счет накопления оперативных данных, так и обновления ядерного массива химических соединений, периодически импортируемого из регистра химических соединений хранилища данных.

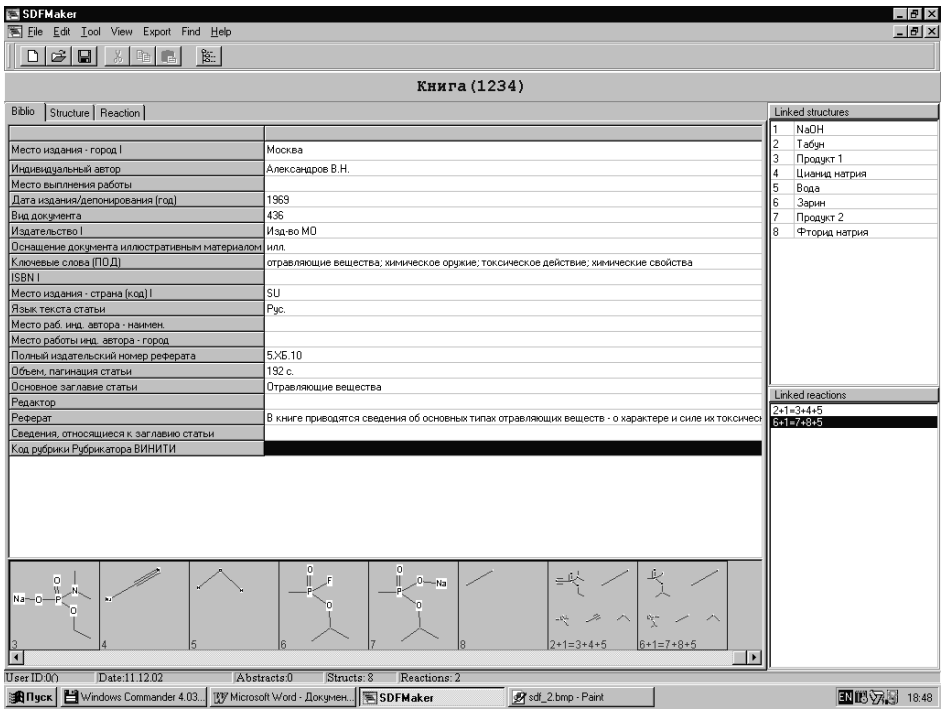

Рис. 3. *Экранная форма АРМ для работы с реферативно-библиографическими данными документа, обрабатываемого в ОТБД*

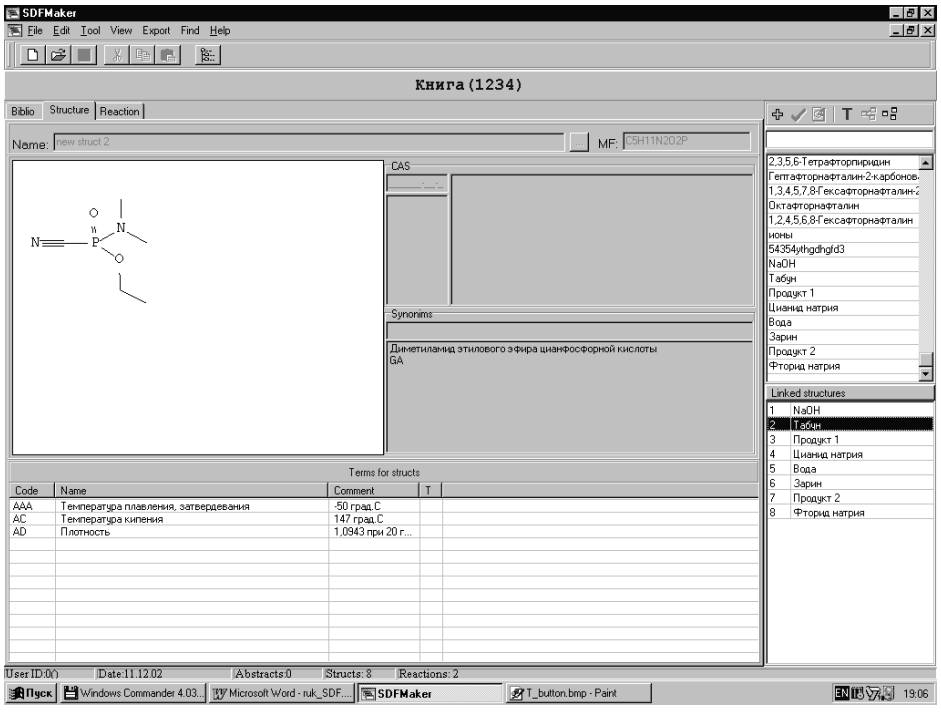

Рис. 4. *Экранная форма АРМ для формализованного представления химических соединений и релевантной фактографии, содержащихся в документах, обрабатываемых в ОТБД*

На рис. 5 приведена экранная форма АРМ для работы с данными по химическим реакциям, включающая возможности:

• просмотр/редактирование уравнения реакции из списка уже формализованных в теку-

щем документе (левое верхнее окно), в том числе путем изменения для текущей реакции состава реагентов и продуктов (средние верхние окна) из списка уже формализованных соединений (правое нижнее окно);

• ввод/редактирование записей условий протекания текущей реакции (катализатор, растворитель, другие неиндексируемые соединения, выходы продуктов, температура и давление) текущей реакции (панель слева посередине) из списка уже формализованных в обрабатываемом документе;

• формирование списка унифицированных записей дополнительных термов свойств (левое нижнее окно) текущей реакции из списка уже формализованных в обрабатываемом документе.

Унификация записей термов свойств химических соединений и реакций, катализаторов, растворителей, других неиндексируемых соединений достигается обязательным использованием встроенных справочников данных по соответствующим информационным объектам.

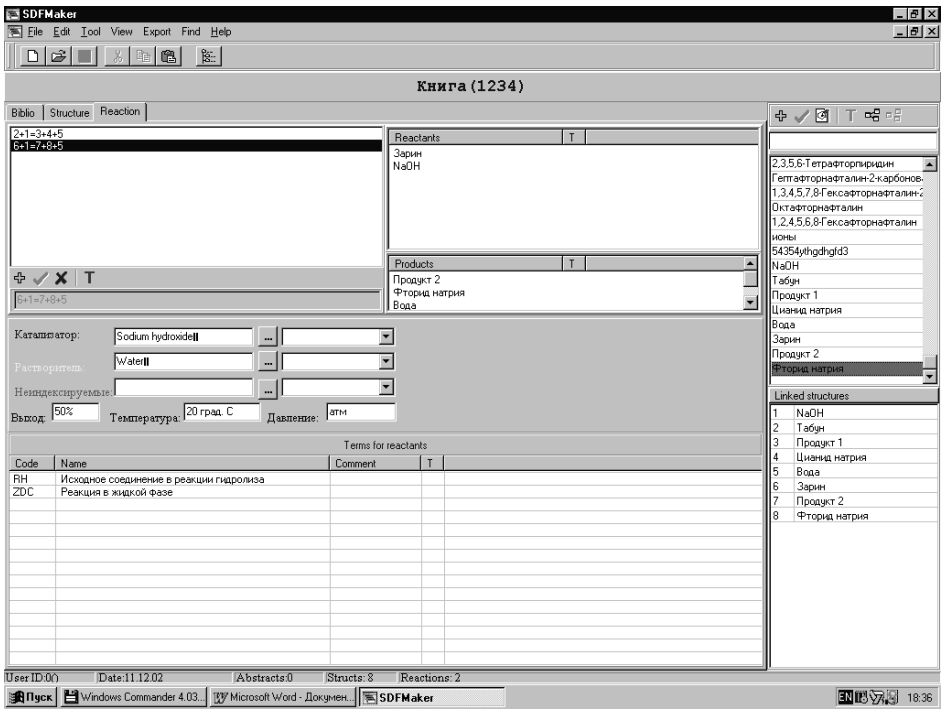

Рис. 5. *Экранная форма АРМ для формализованного представления химических реакций и релевантной фактографии, содержащихся в документах, обрабатываемых в ОТБД*

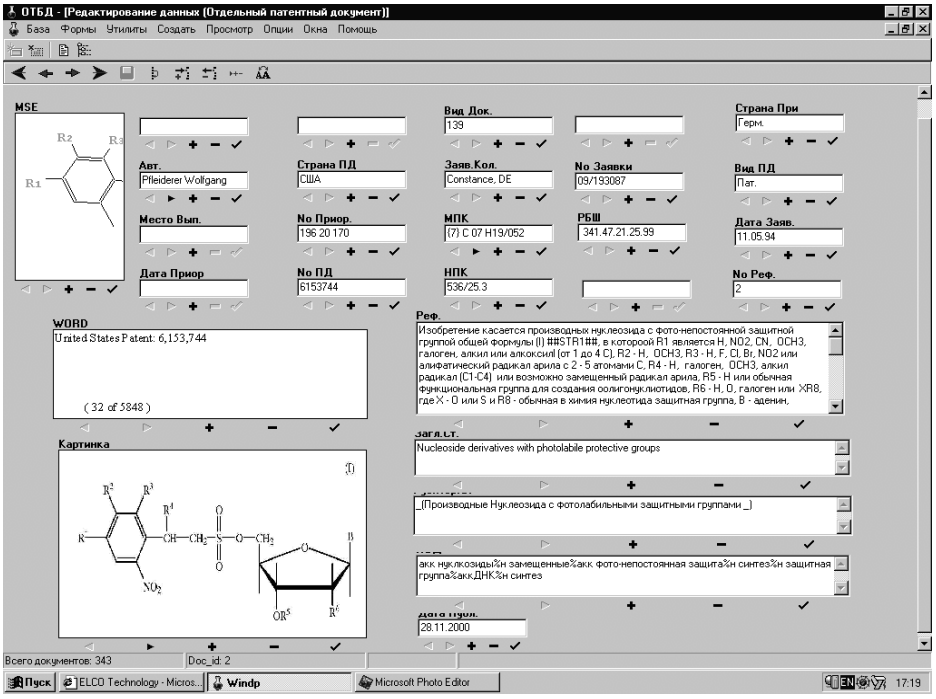

Рис. 6. *Экранная форма АРМ ОТБД для импорта и обработки документов (патентов)*

На рис. 6 приведена экранная форма АРМ ОТБД для импорта и обработки патентных документов. Приведенный пример демонстрирует возможности электронной технологии для экстракции, формализации и содержательной переработки текстовой и графической компоненты сканированного патента, включая: перевод основного названия, составление и набивку реферата и ключевых слов, индексацию химических

 © ВИНИТИ, 2002

соединений и реакций. При этом химико-структурные данные могут быть непосредственно включены как клише в оригинал-макет формируемого в рамках ОТБД реферативного издания.

Выходные оперативные информационные ресурсы, экспортируемые из ОТБД по окончании технологического цикла переработки входного потока документов, включают:

- файл реферативно-библиографических данных текущего выпуска реферативного издания для загрузки в БД хранилища данных;
- файлы химических структур и реакций текущего выпуска реферативного издания для загрузки в БД хранилища данных, в том числе для формирования регистров химических соединений и реакций;
- файлы клише химических структур, реакций, схем и др.;
- файлы оригинал-макетов реферативно-библиографических изданий ВИНИТИ;
- файлы оригинал-макетов дополнительных указателей к реферативным изданиям;
- электронные/бумажные формуляры для удаленной дополнительной обработки документов.

Таким образом, программно-технологические средства ОТБД поддерживают весь комплекс работ по формированию оперативного информационного ресурса, включая подготовку электронных и печатных изданий, а также файлы структурированной информации для загрузки в БД хранилища данных ВИНИТИ.

3. АРМ для регистрации химических соединений и реакций и формирования хранилища данных в области химии и химической технологии

В настоящее время основные информационные ресурсы ВИНИТИ в области химии и химической технологии базируются на последовательном и аддитивном объединении, с одной стороны, реферативно-библиографических данных и, с другой стороны, химико-структурных данных обрабатываемых документов. При этом консолидация химико-структурных данных (химических структур) и реферативно-библиографических данных и возможность их совместного использования обеспечивается ВИНИТИ только в рамках специально формируемых полугодовых выпусков электронных формульных указателей [16], а доступ к данным по химическим реакциям вообще отсутствует. В то же время основная часть указанного ресурса ВИНИТИ успешно адаптируется зарубежными фирмами для полномасштабного коммерческого использования, в том числе в рамках интернет-доступа к БД по химическим соединениям и реакциям [17, 18]. В связи со сложившейся ситуацией в области национальных информационных ресурсов и важностью федерального русскоязычного информационного ресурса в области химии и химической технологии для обеспечения работ по приоритетным направлениям науки и техники актуальным является создание соответствующих корпоративных хранилищ данных ВИНИТИ и других организаций, отвечающих современным требованиям и стандартам на предоставление научной информации.

В соответствии с концепцией АСИО хранилище данных собирает на постоянное хранение преобразованные (формализованные, стандартизированные и унифицированные) данные и/или метаданные [1–7]. При этом БД хранилища представляет систему реестров фактов, интегрированных с системой классификаторов-справочников (тезаурусов, указателей, регистров химических информационных объектов и др.). Структура БД хранилища данных, ее реестров, классификаторов и репозитория, в отличие от операционно-технологической БД (ОТБД), преимущественно ориентирована на обеспечение оптимального доступа к данным и проведение информационно-аналитических исследований.

Химические соединения и реакции являются химическими информационными объектами со сложной семантической структурой, и создание для них реестров фактов предполагает наличие в рамках репозитория метаданных достаточно развитой системы указателей и регистров-классификаторов этих информационных объектов. В соответствии с этим в таблицах учета фактов реестров наряду с полями данных, их определений/размерностей присутствуют поля ссылки (указатели) на соответствующие записи регистров химических соединений и реакций.

© ВИНИТИ, 2002

3.1. Формирование регистра химических соединений и загрузка фактографических данных

Исходными данными для формирования фактографического информационного ресурса хранилища данных по химическим соединениям являются оперативные данные, экспортируемые из ОТБД по окончании технологического цикла переработки входного потока документов, представленные SDF-файлами, включающими последовательные записи химических структур (MOL-представление) с набором полей релевантных данных.

Основные операции, реализуемые при загрузке данных по химическим соединениям в рамках разрабатываемых АРМ пользовательского интерфейса реляционной MS SQL-БД, включают:

- поиск соединений по идентификационным кодам CAS;
- вычисление хэш-индексов химических структур;
- выделение и подсчет структурных дескрипторов («отпечатки пальцев»);
- подструктурный поиск по химическим структурам;
- присваивание уникального регистрационного кода;
- регистрацию данных химического соединения, если необходимо;
- внесение записей в реестры учета фактов-данных для химического соединения, если необходимо.

В режиме пакетной загрузки данных, контролируемой экспертом, АРМ позволяет:

- выбрать список файлов для пакетной регистрации;
- задать степень автоматизации процесса загрузки (от полностью ручного режима до полностью автоматического);
- осуществлять контроль и редактирование, если необходимо, всех полей данных (рис. 7 и 8);
- принимать решения, связанные с погрузкой данных (регистрация нового соединения, обновление данных для текущего соединения, отказ от загрузки данных по текущему соединению, замена только химической структуры и др.).

На рис. 7 и 8 приведены экранные формы АРМ для принятия решения по текстовым полям фактографических данных и химическим структурам, соответственно, при загрузке данных в БД хранилища из SDF-файлов.

В результате загрузки SDF-файлов формируются таблицы регистров (классификаторов/указателей) и реестров учета фактов (данных):

- регистры: молекулярных фрагментов (дескрипторов), молекулярных структур (связанных графов) и химических соединений;
- реестры: «структуры – дескрипторы», «химические соединения – дескрипторы», «химические соединения – отдельные структуры», «химические соединения – общие данные (синонимы, обозначения, индексаторы, библиография)» и «химические соединения – термы свойств».

3.2. Формирование регистра химических реакций и загрузка фактографических данных

Исходными данными для формирования фактографического информационного ресурса хранилища данных по химическим реакциям служат RDF-файлы, содержащие последовательные записи: молекулярных данных реагентов и продуктов с набором полей релевантных фактографических данных – термов свойств.

В рамках разрабатываемых АРМ основные операции, реализуемые при загрузке данных по химическим реакциям, включают:

- выделение наборов химических соединений, входящих в реакции;
- поиск и/или регистрацию каждого химического соединения из выделенных наборов (см. выше);
- регистрацию реакций с присваиванием соответствующих уникальных регистрационных кодов;
- учет фактов/данных по реакциям.

В режиме пакетной загрузки данных, контролируемой экспертом, АРМ позволяет:

- выбрать список файлов для пакетной регистрации;

 © ВИНИТИ, 2002

• задать степень автоматизации процесса загрузки (от полностью ручного режима до полностью автоматического);

• осуществлять контроль и редактирование, если необходимо, всех полей данных.

В результате загрузки RDF-файлов формируются таблицы регистров (классификаторов/указателей) и реестров учета фактов (данных):

• регистры: молекулярных фрагментов (дескрипторов), молекулярных структур (связанных графов), химических соединений и химических реакций;

• реестры: «структуры – дескрипторы», «химические соединения – дескрипторы», «химические реакции – дескрипторы», «химические соединения – отдельные структуры», «химические реакции – отдельные структуры», «химические реакции – химические соединения»; «химические реакции – общие данные (синонимы, обозначения, индексаторы, библиография)» и «химические реакции – термы свойств».

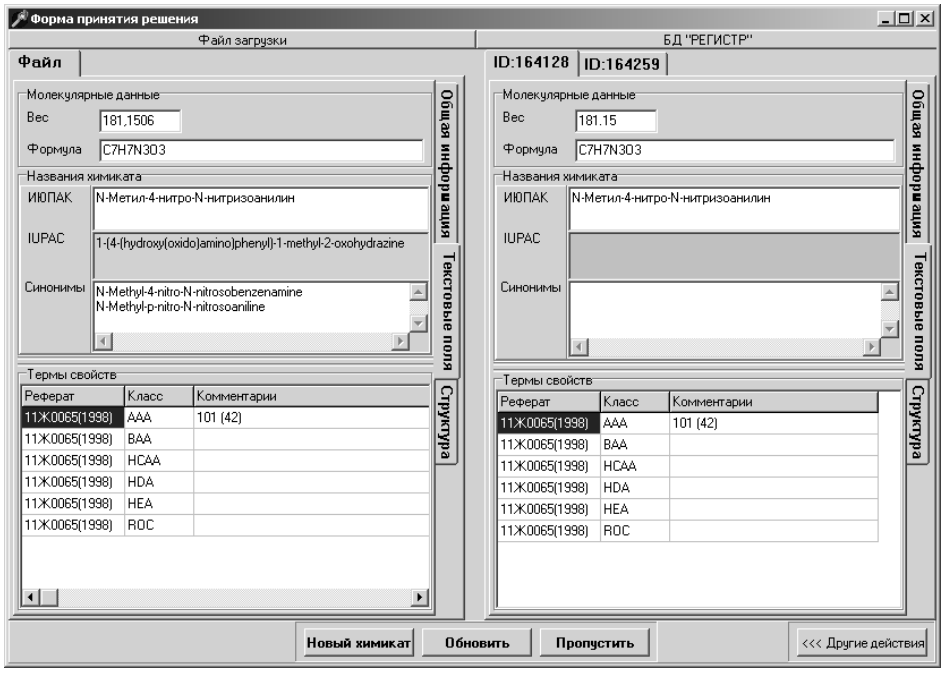

Рис. 7. *Загрузка данных по химическим соединениям в хранилище данных. Экранная форма АРМ, вкладка «Текстовые поля»*

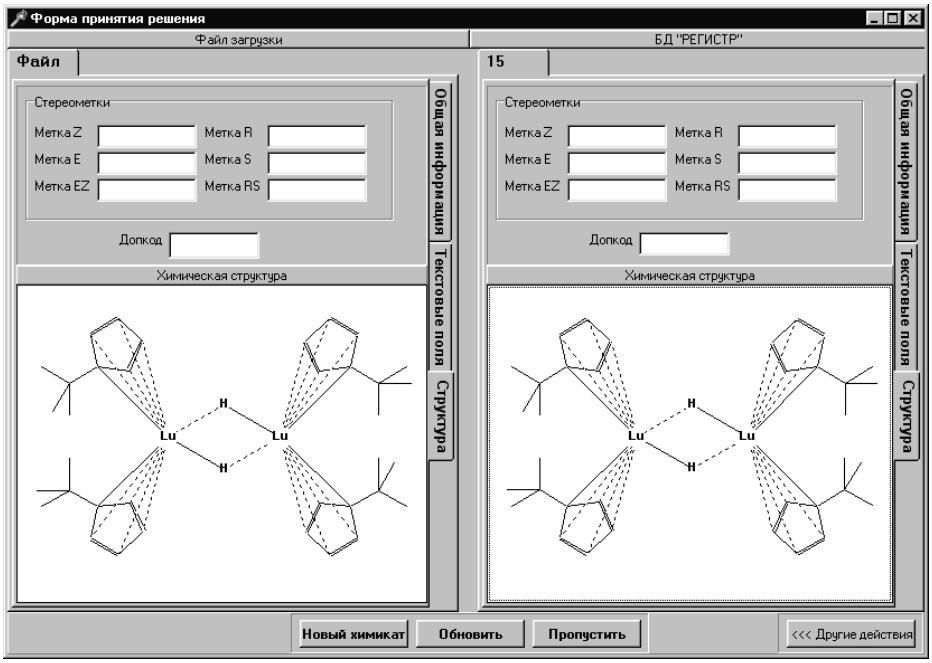

Рис. 8. *Загрузка данных по химическим соединениям в хранилище данных. Экранная форма АРМ, вкладка «Структура»*

Следует подчеркнуть, что консолидация в рамках хранилища химико-структурных и реферативно-библиографических данных осуществляется, с одной стороны, за счет введения соответствующего атрибута в записи реестров фактов/данных по химическим соединениям и реакциям, что предопределяет необходимость наличия в БД хранилища указателя документов – источников информации, в том числе и перерабатываемых в ВИНИТИ. С другой стороны, ведение и для них соответствующих реестров фактов/данных обеспечивает хранение и поиск документальных данных, включая клише струк-

турных формул и/или ссылки на химические соединения и реакции, в том числе зарегистрированные в БД хранилища данных ВИНИТИ и/или других отечественных и зарубежных БД. При этом важно отметить, что способ включения классификаторов – указателей фактов/данных предусматривает возможность их развития как соответствующих разделов репозитория метаданных хранилища данных, в том числе и в плане его интеграции с мировыми информационными ресурсами.

4. АРМ для работы с химико-структурными данными хранилища данных в области химии и химической технологии

Структура БД хранилища данных и программные средства пользовательского интерфейса преимущественно ориентированы на обеспечение оптимального доступа к записям реестров фактов/данных с использованием системы регистров – указателей/классификаторов данных. Разработанные в рамках пользовательского интерфейса макеты АРМ хранилища предназначены для решения типовых задач обеспечения работы с химико-структурными данными:

- просмотр текстовых и графических данных по химическим соединениям и реакциям, включая название, синонимы, химические структуры, поля регистрационных, реферативно-библиографических, фактографических и административных данных;
- поиск/отбор химических соединений и реакций по реферативно-библиографическим данным;
- поиск/отбор химических соединений по текстовым полям: регистрационные коды (ВИНИТИ, CAS, BEILSTEIN и др.), систематическое название IUPAC, молекулярная формула, синонимы названий;
- поиск, отбор и/или классификация химических соединений и реакций по полям соответствующих реестров фактов/данных;
- поиск/отбор химических соединений по химической структуре на точное совпадение и структурно-подструктурное соответствие, в том числе и по Маркуш-формулам;
- поиск/классификация (например, патентная классификация) заданной химической структуры на соответствие Маркуш-формуле обобщенного соединения, зарегистрированного в БД;
- поиск/отбор химических реакций, в которых принимает участие заданное химическое соединение, в том числе как реагент (прекурсор) и/или продукт реакции.

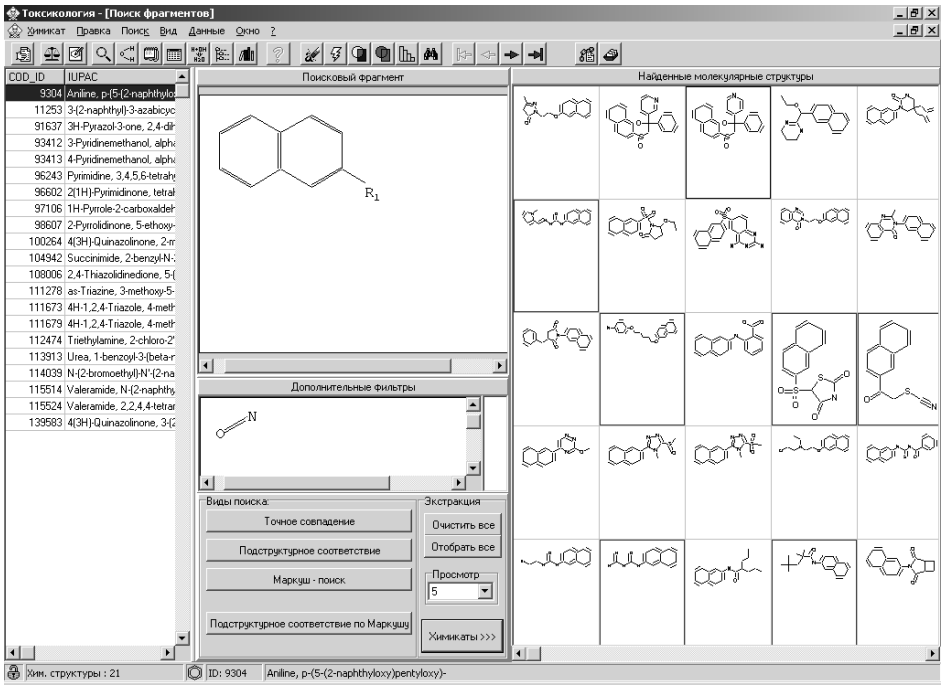

Рис. 9. *Экранная форма АРМ для поиска химических соединений по молекулярным данным*

На рис. 9 в качестве примера приведена экранная форма АРМ для поиска химических соединений по молекулярным данным (химической структуре, дескрипторам, фрагментам, формулам Маркуша). Особенность программно-технологического решения для работы с химико-структурными данными состоит в том, что в этом случае регистр химических

 © ВИНИТИ, 2002

структур выступает в качестве указателя для «подходящих» химических соединений, а молекулярные дескрипторы выступают уже в качестве соответствующих записей реестра фактов для химических структур.

Сходная схема организации взаимодействия «регистр → реестр/регистр → реестр/регистр → ….» для доступа и работы с данными хранилища реализована при разработке и других АРМ. На рис. 10 в качестве примера приведена экранная форма АРМ для поиска, ввода и редактирования химических реакций и релевантной фактографии. В этом случае регистр химических соединений выступает в качестве указателя «подходящих» химических реакций. В то же время соответствующие молекулярные данные, выступающие в качестве данных о химических соединениях, интерпретируются как метаданные (данные о данных).

Следует подчеркнуть, что наряду с решением типовых задач обеспечения работы с химико-структурными данными, в рамках разрабатываемых АРМ хранилища данных предусмотрена возможность дополнительной обработки накапливаемой информации, в том числе и программно-технологической реализации поддержки принятия решения на основе специального структурирования, агрегирования и моделирования фактов/данных.

На рис. 11 в качестве примера приведена экранная форма АРМ для формализации количественных данных с использованием реестров

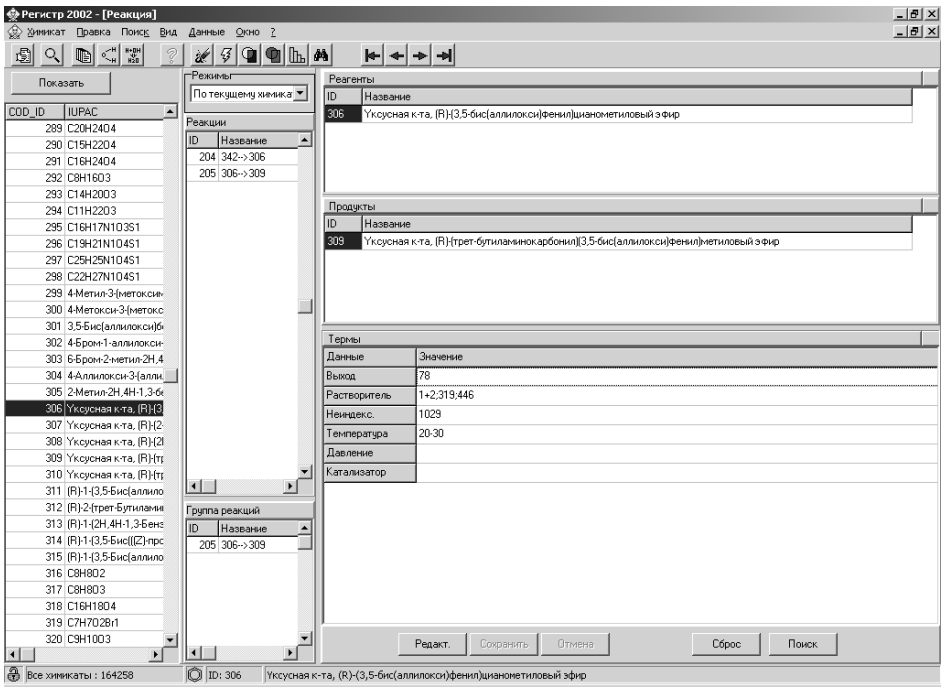

Рис. 10. *Экранная форма АРМ для поиска, ввода и редактирования химических реакций по соединениям (реагенты и/или продукты реакции) и релевантной фактографии*

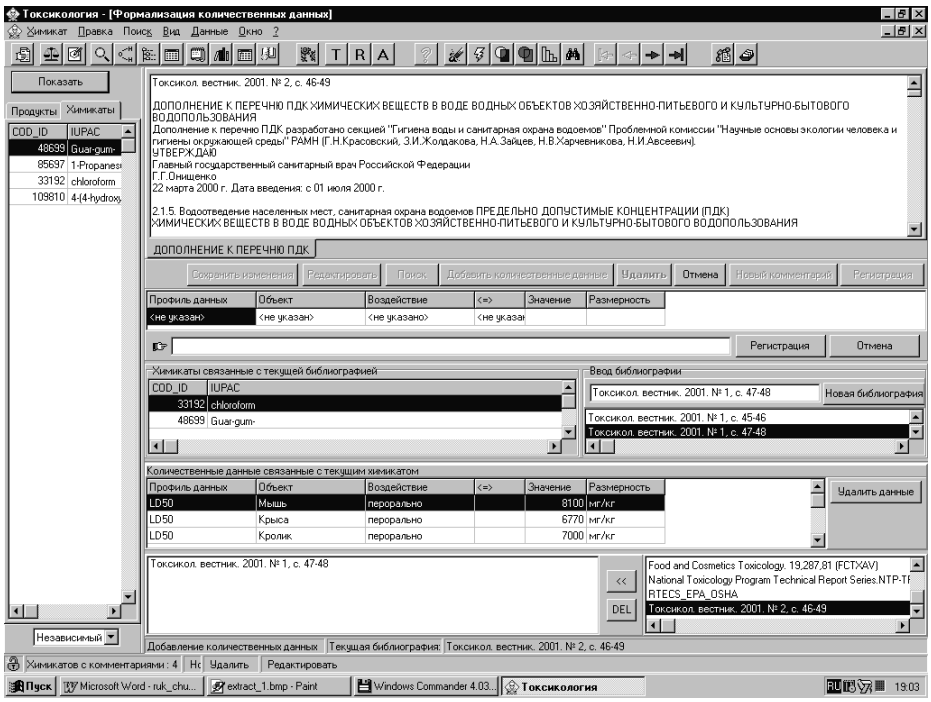

Рис. 11. *Экранная форма АРМ для формализации количественных данных реестра фактов/термов свойств химических соединений*

учета фактов/термов свойств химических соединений, релевантных реферативно-библиографических данных, указателей-классификаторов количественных данных («Профили», «Объекты», «Воздействия», «Размерности») и других метаданных репозитория хранилища данных. В этом плане важно подчеркнуть, что метаданные репозитория обеспечивают централизованное управление семантическими объектами информационной инфраструктуры БД хранилища, позволяют автоматизировать поиск, формализацию и унификацию данных, в том числе в рамках запросов/процедур, сформулированных на естественных языках. При этом допускается использование внешних информационных ресурсов (например, электронных справочников, классификаторов и БД) и программных средств (например, программ квантово-химических расчетов, программ восстановления химических структур по названиям и наоборот – систематического названия по химической структуре).

На рис. 12 приведена принципиальная схема информационной и статистической связи данных, реализованная в БД хранилища для решения задач поиска, классификации и прогноза свойств химических веществ. Система информационно-статистических моделей включает:

- модели количественных и категорированных данных;
- информационные модели связи данных;
- статистические модели классификации химических соединений (Байесовская классификация);
- межтабличные статистические модели связи данных («химическая структура – свойство», «свойство – свойство»).

5. Программно-технологические средства формализованного представления сведений о химическом производстве

В рамках разрабатываемой концепции формирования хранилища данных в области химии и химической технологии в качестве базовых элементов описания химического производства выступают аппараты, технологические процессы, продукты/химические соединения, определяемые в соответствующих указателях – регистрах/классификаторах. Использованные при разработке программных средств пользовательского интерфейса БД хранилища данных мето-

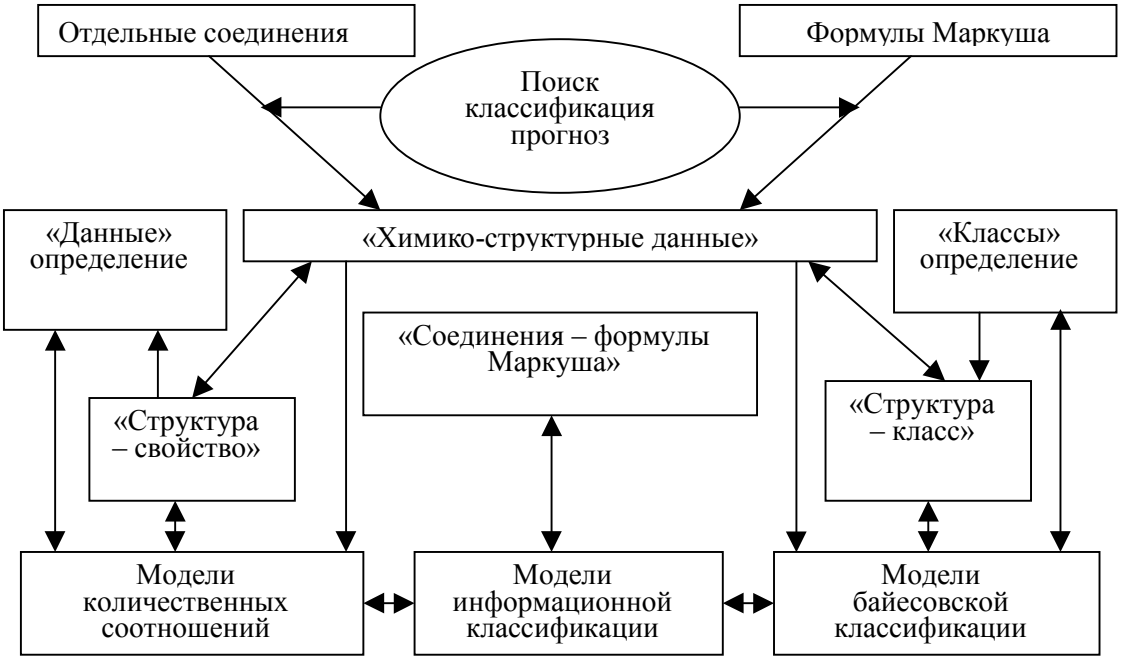

Рис. 12. *Принципиальная схема информационной и статистической связи данных, реализованная в БД хранилища для решения задач поиска, классификации и прогноза свойств химических веществ*

 © ВИНИТИ, 2002

ды анализа и формализации соответствующих технологических схем, так же как и других химических информационных объектов, основаны на применении теории графов. Технологические схемы дают развернутое, в том числе и графическое, представление о химическом производстве. Реализованные в рамках АРМ пользовательского интерфейса БД хранилища данных информационные проекции/экранные формы определения и отображений технологических схем включают:

- производство, производителя, продукты, химические соединения;
- технологические операции и процессы, включая химические реакции;
- элементы конструкции и аппараты.

На рис. 13 в качестве примера приведена экранная форма АРМ для поиска технологических схем. В качестве элементов навигации здесь выступают указатели химических соединений (химикатов)/продуктов – ингредиентов различных типов (растворителей, реагентов, продуктов и др.) и различного типа операций, включая химические реакции. Доступ к данным для выбранной технологической схемы осуществляется в рамках специальной экранной формы (рис. 14), включающей вкладки: «Общая информация», «Процессы и операции», «Аппараты и элементы аппаратов».

Разработанные информационные и программно-технологические средства формализо-

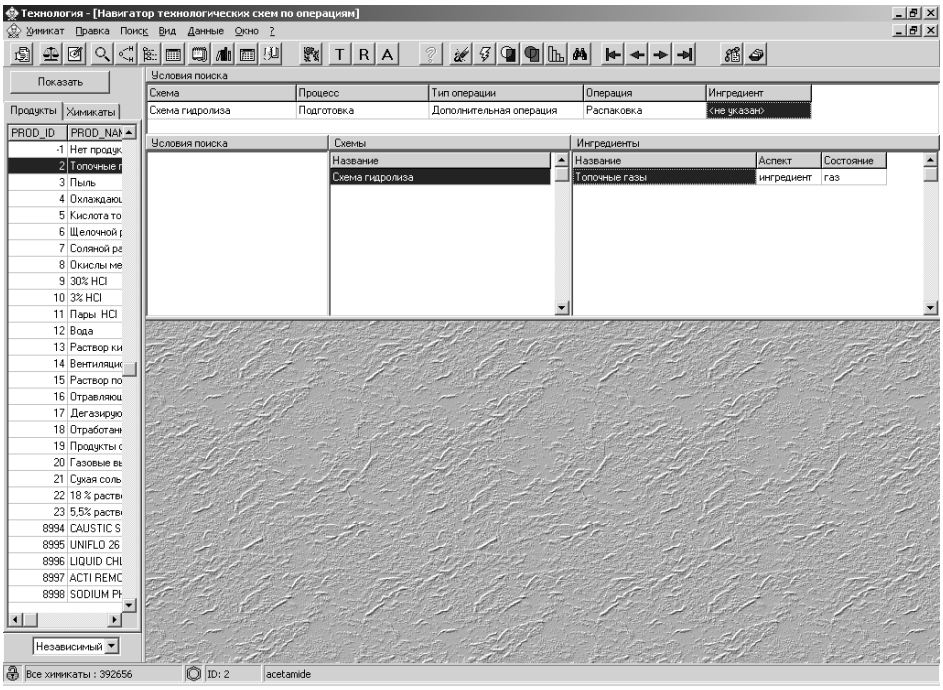

Рис. 13. *Экранная форма АРМ для поиска технологических схем по операциям и химическим ингредиентам (растворителям, реагентам, продуктам и др.)*

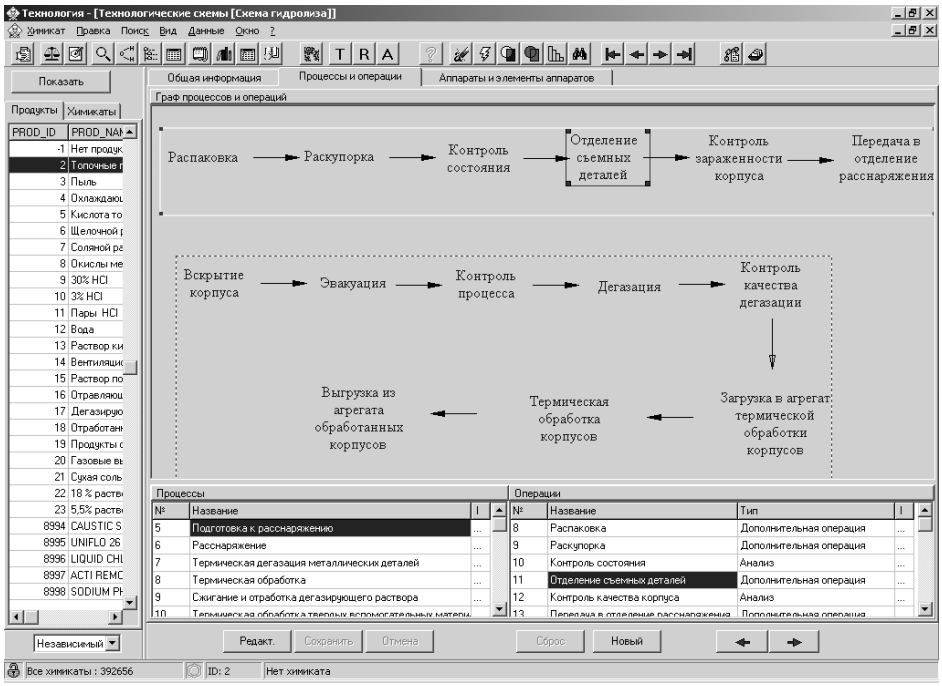

Рис. 14. *Экранная форма АРМ для работы с данными по выбранной технологической схеме химического производства. Вкладка «Процессы и операции»*

ванного представления сведений о химическом производстве в БД хранилища данных отражают начальный этап в разработке технологии создания соответствующего корпоративного информационного ресурса. При этом программно-технологические решения по хранилищу данных рассматриваются как необходимый этап развития ОТБД и АРМ для более полной и содержательной переработки входного потока научно-технической и патентной информации в области химии и химической технологии.

Заключение

Проведенные опытная разработка и развитие программно-технологических решений по регистрам химических соединений и реакций являются важным этапом развития информационных технологий формирования крупных информационных ресурсов, базирующихся на унифицированном представлении, классификации и регистрации химических объектов (от структурных данных до материальной химии и технологии) в документально-фактографической БД хранилища данных. Созданы прототипы регистров химических соединений, реакций и смесей, химико-технологических схем, процессов, аппаратов и аналитических методов. Разработан и подготовлен к внедрению комплекс новых информационных и программно-технологических средств формализованного, в том числе и графического, представления сведений о химическом производстве.

Основной задачей следующего этапа работ является опытная отработка и развитие программно-технологических решений на реальных информационных массивах ВИНИТИ и других производителей/поставщиков информации по химии, химической технологии и смежным областям знаний, в том числе для обеспечения WWW-доступа к информационным ресурсам по химико-структурным данным.

Литература

1. *Афонина Н.Н., Бирюков А.Л., Гончаренко Д.И., Зацепин В.М., Иванченко В.А., Мартиросов А.К., Нефедов О.М., Остапчук Р.В.* Развитие банка данных информационных ресурсов ВИНИТИ по химии и химической технологии: от структурных данных к материальной химии и технологии. 1. Разработка концепции и исследовательских прототипов регистров химических информационных объектов // Крит. технол. Мембраны. 2001. №12, с. 38–51.
2. *Бирюков А.Л., Зацепин В.М., Иванченко В.А., Гончаренко Д.И.* Адаптивная система информационного обеспечения по мембранным технологиям. 1. Основные принципы // Крит. технол. Мембраны. 2000. № 5, с. 29–44.
3. *Бирюков А.Л., Зацепин В.М., Иванченко В.А., Гончаренко Д.И.* Адаптивная система информационного обеспечения по мембранным технологиям. 2. Формирование проблемно-ориентированного информационного ресурса // Крит. технол. Мембраны. 2000. № 6, с. 61–70.
4. *Зацепин В.М., Иванченко В.А.* База данных по химическим веществам «Токсикология» для аналитико-информационных технологий // Химия в России. 1999. № 1, с. 14–17.
5. *Зацепин В.М., Иванченко В.А.* Химическая информатика: системология, состояние, проблемы // Тез. докл. 7-го Междунар. конгресса «Человек и лекарство». М. 2000, с. 498.
6. *Зацепин В.М., Иванченко В.А., Афонина Н.Н., Гончаренко Д.И., Мартиросов А.К., Остапчук Р.В.* Интегральная информационно-аналитическая система по химическим веществам для поддержки биомедицинских и токсикологических исследований // Сб. тез.: Научно-практич. конф. «Медицинские информационные технологии». М., ВВЦ. 2001, с. 23–24.
7. *Зацепин В.М., Иванченко В.А.* Конструктивная модель биоинформатики и стратегия интегральных информационно-моделирующих систем // Тез. докл. 8-го Междунар. конгресса «Человек и лекарство». М. 2001, с. 482.
8. *Сафронова О.С., Бирюков А.Л., Гончаренко Д.И., Зацепин В.М., Иванченко В.А.* Разработка системы формализации и регистрации информационных объектов в интегральной многоаспектной базе данных по химии и химической технологии // Материалы 6-й Междунар. конф. «НТИ-2002», 16–18 октября 2002. М.: ВИНИТИ РАН. 2002, с. 303.
9. http://www.cas.org
10. http://www.daylight.com/dayhtml/smiles
11. http://www.tripos.com
12. http://www.mdli.com
13. http://www.daylight.com
14. http://www.acdlabs.com
15. http://www.cambridgesoft.com
16. *Бирюков А.Л., Буторина Л.С., Дубицкая Н.Ф., Зацепин В.М., Князева Г.Р., Красотченко В.В., Марголин Л.Н., Рахманина А.В., Трепалин С.В.* Компьютерный формульный указатель к РЖ Химия ВИНИТИ // Материалы 4-й Междунар. конф. «НТИ-99», 17–19 марта 1999. М.: ВИНИТИ РАН. 1999, с. 53–54.
17. http://www.infochem.de/spresi.htm
18. http://www.daylight.com/products

 © ВИНИТИ, 2002