ИНФОРМАЦИОННЫЙ ПОИСК

УДК 004.65:54

Ю.Е. Бессонов, Б.С. Фельдман, Н.И. Чуракова, Е.В. Кочетова

Поиск и отображение информации о химических реакциях в базе структурных данных по химии ВИНИТИ РАН

Описана База химических реакций, входящая в состав Базы структурных данных по химии ВИНИТИ РАН. Рассмотрены вопросы визуального отображения информации о реакциях, поиска реакций по их характеристикам. Представлена программа поиска в пользовательской базе данных химических реакций и приведены примеры выполнения поиска. Намечены перспективы дальнейшего развития поисковой системы химических реакций.

Ключевые слова: *база химических реакций, поиск и отображение информации в базах данных по химии*

DOI: 10.36535/0548-0027-2022-03-2

ВВЕДЕНИЕ

Специализированная база данных структурной химической информации ВИНИТИ РАН (далее База СД) – крупнейшее в России хранилище структурных химических данных, содержащее информацию о более чем 7,2 млн химических структур, 4,1 млн химических реакций и 15,2 млн свойств химических соединений. База СД включает базу структур химических соединений и базу реакций, в которых эти соединения принимают участие.

В настоящее время пополнение Базы СД производится с помощью программного комплекса CBASE32 [1], являющегося развитием первой российской программы графической обработки структурных данных по химии CBASE16 [2]. CBASE32 обеспечивает ввод и обработку структурных, фактографических и библиографических данных первоисточников и состоит из компонентов:

- программное обеспечение для ввода информации о соединениях;
- программное обеспечение для ввода информации о реакциях;
- электронный справочник химических соединений;
- электронный справочник именных реакций.

Информация о химических соединениях содержит молекулярную формулу, систематическое и/или тривиальное название, химическую структуру, международный химический идентификатор IUPAC InChI и его хешированное символьное представление фиксированной длины InChIKey, а также предметные характеристики соединения (физико-химические свойства, сведения о реакционной способности, биологической активности, токсикологии, применении и др.).

Информация о химических реакциях – это сведения об участниках реакции (реактанты, растворители, катализаторы, прочие участники реакции), условиях реакции (температура, давление, время), о выходе, а также предметные характеристики, дающие дополнительную текстовую информацию о реакции (теоретические, физико-химические и технологические аспекты изучения реакций**,** специальные методы синтеза**,** области применения реакций и т.д.). В случае многостадийных реакций приводятся данные, описывающие протекание реакций по отдельным стадиям.

Электронный справочник химических соединений (далее Глоссарий) – структурированная и пополняемая база химических соединений [3, 4]. Он позволяет стандартизировать представление структурных и фактографических данных о химических соединениях и упрощать процедуру ввода информации о соединениях и реакциях в Базу СД, а также осуществлять поиск соединения по молекулярной формуле, структуре и названию. Если соединение найдено в Глоссарии, то непосредственно из него оно вводится в Базу СД со всеми атрибутами и при этом получает номер, под которым оно фигурирует в Глоссарии. Пополнение и редактирование Глоссария осуществляется средствами CBASE32. В настоящее время Глоссарий содержит более 3000 соединений.

Электронный справочник именных реакций – структурированная и пополняемая база именных реакций, содержащая более 300 именных реакций в русском и англоязычном варианте, в которой возмо-

жен поиск по записям названий реакций. Справочник был сформирован в результате обработки данных из Интернета и литературных источников [5–7].

База структурных данных содержит релевантные, критически оцененные, тщательно отобранные и проанализированные высококвалифицированными специалистами-химиками данные, что является одним из ее серьезных преимуществ на фоне иных разнообразных потоков химической информации в современном информационном мире.

Когда накоплен большой массив востребованной информации, в нем необходимо обеспечить эффективный поиск, под которым понимается быстрое выполнение запросов, точность, полнота и наглядность результатов. Для поиска информации о химических соединениях, представленных в Базе СД, были разработаны эффективные системы, работающие в интерактивном [8, 9] и автономном режимах [10, 11]. Система поиска в автономном режиме также дает возможность просматривать сведения о химических реакциях. Однако переход к данным о реакциях в этом режиме возможен только через первоначальный поиск по химической структуре, являющейся ее участником. Поиск в этом случае выполняется от соединения к реакции (соединение → реакция).

В настоящей статье рассматривается система поиска химической информации, выполняемого в противоположном направлении: от реакции к участвующим в ней соединениям (реакция → соединение). Предлагаемая система позволяет эффективно выполнять запросы, касающиеся непосредственно химических реакций, что является существенным дополнением к системе поиска в автономном режиме.

Особенность предлагаемой системы поиска реакций – ее функционирование в автономном режиме на локальных компьютерах. При этом данные о химических реакциях хранятся в локальных БД ограниченного объема, получаемых из Базы структурных данных. Объем локальных БД соответствует календарному периоду времени формирования массива Базы СД: месяц, квартал, полугодие, год.

СТРУКТУРА БАЗЫ СД. ФАЙЛЫ ОБМЕННЫХ ФОРМАТОВ

Массив Базы структурных данных систематизирован по годам. Каждый год включает 12 выпусков (номеров) – наборов библиотек исполнителей[1]. *Библиотека исполнителя* – это структурная единица пономерного массива Базы СД, создаваемая одним исполнителем и содержащая информацию о химических соединениях и реакциях, которая получена на основе обработки определенного набора первоисточников. Для удобства работы исполнитель может иметь в номере массива Базы СД несколько библиотек, обозначаемых заглавными буквами латинского алфавита (A, B, C и т.д.). На физическом уровне Библиотека исполнителя представляет собой бинарный файл, содержащийся в каталоге соответствующего номера.

Таким образом, файловая система Базы СД имеет следующую структуру:

<Имя диска[2]>\ <Год>\ <Номер >\<Библиотека>

В программном комплексе CBASE32 имеется сервис, позволяющий выгружать данные из Базы СД в текстовые файлы обменных форматов для структур (SDF, InChI, Smiles) и для реакций (RDF). Система поиска информации о химических реакциях, представляемая в нашей статье, работает с данными, которые получаются в результате обработки rdf- и sdf-файлов.

Формат реакций и данных RDF (reaction-data file), созданный для хранения информации о химических реакциях, был разработан и опубликован компанией Molecular Design Limited (MDL) [12]. Он содержит набор записей RXN химических реакций. Реакционный формат RXN служит для кодирования химических реакций путем указания их реактантов, продуктов и других участников.

Выгрузка обменных файлов из Базы структурных данных дает два набора файлов форматов RDF и SDF[3], причем между файлами rdf и sdf имеется взаимно-однозначное соответствие: <имя файла>. rdf ↔ <имя файла>. sdf. Выгруженные файлы помещаются в каталоги в соответствии с файловой структурой Базы СД: <Год>\<Номер>\<Файл>. Данные Глоссария выгружаются в отдельный файл gloss.sdf.

Имена файлов – имена Библиотек вида

nnnnWmmXyyyy,

где nnnn – четырехзначный код исполнителя, создавшего файл, W – обозначение библиотеки исполнителя (W=A,B,C,D, …), mm – цифровое обозначение номера в рассматриваемом номере массива Базы СД, X – разделитель, yyyy – обозначение года.

В rdf-файле содержится информация о реакциях и ссылки на описания всех участников реакций в соответствующем файле sdf или в Глоссарии (файл gloss.sdf). Согласно технологическому режиму создания Базы структурных данных используются два вида ссылок на соединения: номера записей структур в одноименных sdf-файлах (это технические номера соединений, соответствующие их порядковым номерам при вводе в Базу СД) и номера записей структур, соответствующие их номеру в Глоссарии [3]. Ссылка на структуры в Глоссарии отмечается символом «*» перед номером соединения.

В rdf-файле приведена следующая информация о химических реакциях:

- порядковый номер записи реакции в файле;
- символическая запись уравнения реакции;
- число реактантов и количество продуктов;
- стадия, в которой участвует каждый реактант;
- выход продуктов в процентах;

[1] Исполнитель – это эксперт-химик, выполняющий ввод информации в Базу СД.

[2] Логическое имя устройства внешней памяти.

[3] Для обозначения формата содержимого файла используются прописные буквы (RDF, SDF), а строчные буквы применяются для обозначения расширения соответствующего файла.

- условия проведения реакции на каждой стадии (температура, время, давление);
- участники реакции (реактанты, продукты, катализаторы, растворители и др.) с указанием их названий, а также ссылки на соответствующие ресурсы;
- предметные характеристики реакций, а также предметные характеристики, относящиеся к реактантам и продуктам.

ПОИСК И ОТОБРАЖЕНИЕ ИНФОРМАЦИИ О ХИМИЧЕСКИХ РЕАКЦИЯХ

Общие требования к системе поиска химических реакций

При разработке современных информационно-поисковых систем, как правило, руководствуются следующими требованиями:

а) быстрый поиск;

б) удобный интерфейс для наглядного и полного отображения результатов поиска;

в) полнота информации, содержащейся в результатах поиска.

Требование а) создает проблемы алгоритмического характера в случае больших баз данных. Особенно это характерно для структурного поиска химической информации. Достаточно эффективно справиться с такой задачей позволяет иерархическая организация данных, которая будет описана ниже.

Соблюдение требований б) и в) из-за большого числа атрибутов химических реакций вызывает трудности, обусловленные ограниченностью площади экрана компьютера, особенно если выводится структурная информация в графическом виде. Для их решения был применен принцип, который можно назвать «движение от общего к частному». Клиент сначала видит общую картину представленной в локальной базе данных информации. Затем он может поэтапно детализировать такую картину вплоть до показа значений отдельных атрибутов химических реакций. Этому способствует иерархическая организация данных – после того как реакция найдена, все ее атрибуты возможно получить без дополнительного поиска. Аналогичный подход был ранее использован при создании автономной системы структурного поиска в Базе структурных данных [10].

Структурирование данных о химических реакциях

Организация данных о реакциях в виде наборов бинарных файлов Базы СД, а также в виде rdf- и sdf-файлов, распределенных по каталогам, не позволяет вести эффективный поиск необходимой информации. Требованию быстрого поиска должна отвечать база данных химических реакций, построенная по иерархическому принципу. Для решения такой задачи информацию о каждой реакции, записанной в rdf-файле**,** предлагается дополнить поисковыми образами и представить в упорядоченном виде, как это показано на рис. 1.

Пояснения и примеры

Общие атрибуты реакции:

- имя rdf-файла (например, 9015B01X2019.rdf);
- порядковый номер записи реакции в данном файле rdf;
- *сигнатура* уравнения реакции – упорядоченная пара чисел (r, p), где r – число реактантов, а p – число продуктов;
- число стадий реакции.

Атрибуты стадии реакции:

- номер стадии;
- *условия протекания реакции* – информация о температуре, давлении, времени и прочих условиях;
- число реактантов, растворителей, катализаторов и прочих участников на данной стадии.

Предметные характеристики реакций представляют текстовую запись, дающую сведения о классе реакции, специальных методах получения веществ, различных аспектах изучения и применения реакций, а также о дополнительных условиях их проведения.

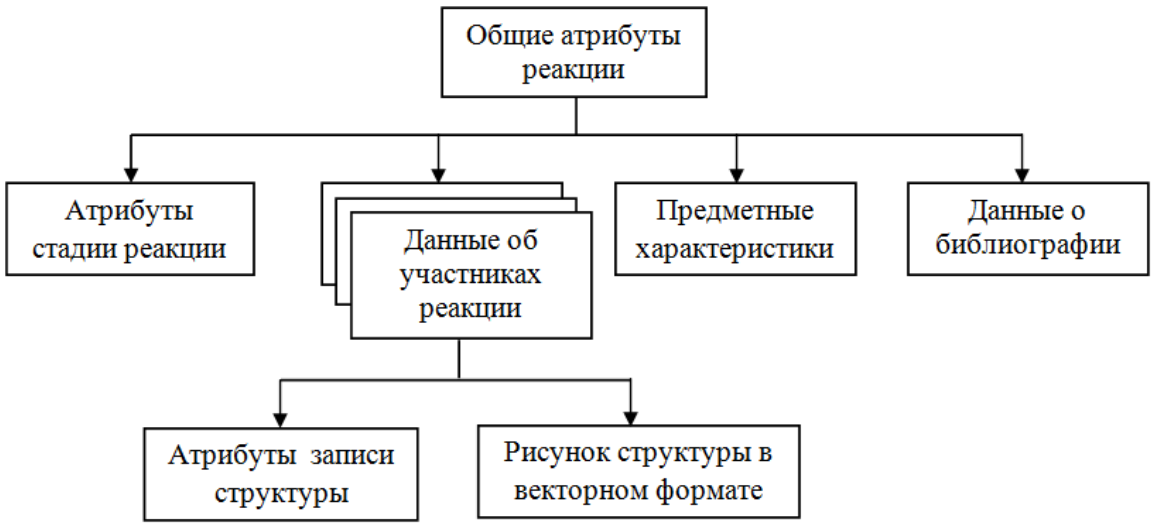

Рис. 1. Иерархическая организация данных о химической реакции.

Таблица

Пример записей полей в формате SDF

Заголовок поля	Описание	Примеры строк содержания
> <RN>	Номер документа	01X0310(2012)
> <LIB>	Номер библиотеки	9301B01X2012
> <SID2>	Код источника библиографии	J09811432216 Системный идентификатор документа-статьи в ВИНИТИ РАН
> <IX>	Код индексатора	9301
> <IT>	Код оператора (вводчика)	9301
> <II>	Номер структуры в документе	12
> <MW>	Молекулярный вес	714.58
> <BF>	Молекулярная формула	C36H37Mo1N4P3
> <SN>	Систематическое название	Бис(эта-диазот)(P,P,P',P",P"-пентафенилдиэтилентрифосфин-P,P',P")(эта-этен)молибден
> <TERMS>	Коды предметных характеристик	BAA HCA$ (13)C, (31)P HCAA HDA RA

К атрибутам участника реакции относятся:

- *роль участника реакции* – одно из следующих значений: Реактант, Продукт, Растворитель, Катализатор, Прочий участник;
- номер структуры в документе sdf-файла или в Глоссарии (если он есть);
- текстовый идентификатор структуры (если ее номер отсутствует);
- молекулярная формула;
- выход в процентах (для продуктов реакции);
- идентификатор источника записи структуры (sdf-файл или Глоссарий).

Рисунок структуры в векторном формате формируется из записей структур в файлах sdf с помощью процедуры графического модуля программы PSDF2PIC (Бессонов Ю.Е. Свидетельство № 2016616626 о государственной регистрации программы PSDF2PIC от 16 июня 2016г. и [13]), входящей в Конвертор данных (см. следующий раздел).

Атрибуты записи структуры – это информация о химической структуре, полученная на основе sdf-файла, содержащая стандартные записи в формате SDF, определяющие структуру молекулы и особенности атомов (блок атомов, блок связей и др.), а также дополнительные записи, учитывающие специфику формирования Базы СД – номер документа, номер библиотеки, номер индексатора и др. Каждая дополнительная запись характеризуется строкой заголовка и следующими за ней строками содержания. В таблице приведены виды дополнительных записей с пояснениями и примерами.

Конвертор данных

Для создания базы данных реакций, построенной по иерархическому принципу, был разработан комплекс программ, называемый *Конвертор данных*. Основная его задача – создание иерархии данных, ориентированной на эффективный поиск и наглядное представление результатов. Поскольку иерархическая база данных реакций, создаваемая Конвертором данных**,** предназначена для предоставления химической информации пользователям, она будет в дальнейшем называться *пользовательской базой данных реакций***.**

Конвертор данных включает модули:

- формирования иерархии данных;
- построения графической информации;
- получения текстовых данных.

Модуль формирования иерархии данных работает следующим образом. Данные из каталога, содержащего rdf- и sdf-файлы, помещаются в оперативную память, где формируются иерархически связанные ссылками массивы записей. Самый верхний уровень – общие данные о реакциях, второй уровень – данные о стадиях, об участниках реакции и о предметных характеристиках, третий уровень – данные о структурах участников, взятые из sdf-файлов (включая gloss.sdf). Затем указанные массивы записываются в бинарный файл General.sdr, содержащий данные первого и второго уровня иерархии.

Модуль построения графической информации формирует файлы, содержащие рисунки химических структур и реакций в векторном формате emf.

Модуль получения текстовых данных создает файлы в текстовом формате с данными о названиях химических соединений, о предметных характеристиках и др., а также библиографическими данными первоисточников.

Набор полученных файлов составляет пользовательскую базу данных реакций. Эта база вместе с предоставляемой пользователям программой поиска химических реакций называется *автономной системой поиска реакций*.

Пользовательские базы данных реакций

Традиционный способ предоставления данных пользователям состоит в том, что клиент получает через Интернет доступ ко всей базе данных, в кото-

рой содержится интересующая его информация. Размер базы данных химических реакций может достигать порядка 10^6 записей как, например, в базах данных Reaxis [14] или в CASReact [15–17].

В отличие от традиционного наш подход основан на концепции пользовательских баз данных, суть которого состоит в следующем. Клиенту предоставляется ограниченный массив данных из Базы СД, сформированный по тем или иным признакам (как правило, это календарный период времени формирования массива: месяц, квартал, полугодие, год). Пример – база, сформированная в результате обработки годового объема выпусков одного конкретного научного журнала.

Пользовательская база данных может быть так же тематической – ориентированной на конкретного заказчика. Как будет показано ниже, работа автономной системы поиска реакций возможна на обычном персональном компьютере в автономном режиме без использования внешних СУБД. Это позволяет системе оперативно предоставлять достаточный объем актуальной химической информации для пользователей, не обладающих большими вычислительными мощностями.

Подобный подход успешно применялся в ВИНИТИ в 1999–2009 гг., когда заказчикам на условиях подписки периодически поставлялись формульные указатели к выпускам реферативного журнала «Химия», содержащие информацию о химических структурах. Подход также применяется в настоящее время при создании локальных баз данных для системы автономного поиска химических структур [10].

ПОИСКОВЫЕ ФУНКЦИИ СИСТЕМЫ И ОСОБЕННОСТИ ОТОБРАЖЕНИЯ РЕЗУЛЬТАТОВ ПОИСКА

Интерфейс главного окна

При старте программы открывается главное окно, интерфейс которого содержит поле для отображения результатов поиска химических реакций (в левой части). При нажатии на кнопку «Читать базу» в этом поле отображается *дерево химических реакций*. Оно имеет два уровня. Интерфейс дерева продемонстрирован на рис. 2, из которого, например, видно, что узел с пометкой «1 2» находится на первом уровне, а подчиненный ему узел с пометкой «C11H18O1=C11H15N1O5+C11H16N2O6» – на втором. Узлы первого уровня соответствуют лексикографически упорядоченным значениям сигнатуры уравнений реакций. Для каждого узла первого уровня подчиненные ему узлы соответствуют уравнениям реакций, в которых реактанты и продукты представлены в виде молекулярных формул. Порядок записи символов в формулах подчиняется правилу Хилла, а узлы одного уровня располагаются сверху вниз в лексикографическом порядке. В этом же порядке располагаются записи членов уравнений в левой (реактанты) и в правой (продукты) частях. Такая организация интерфейса обеспечивает квалифицированному пользователю быструю навигацию по базе данных, позволяя в отдельных случаях не прибегать к запросам.

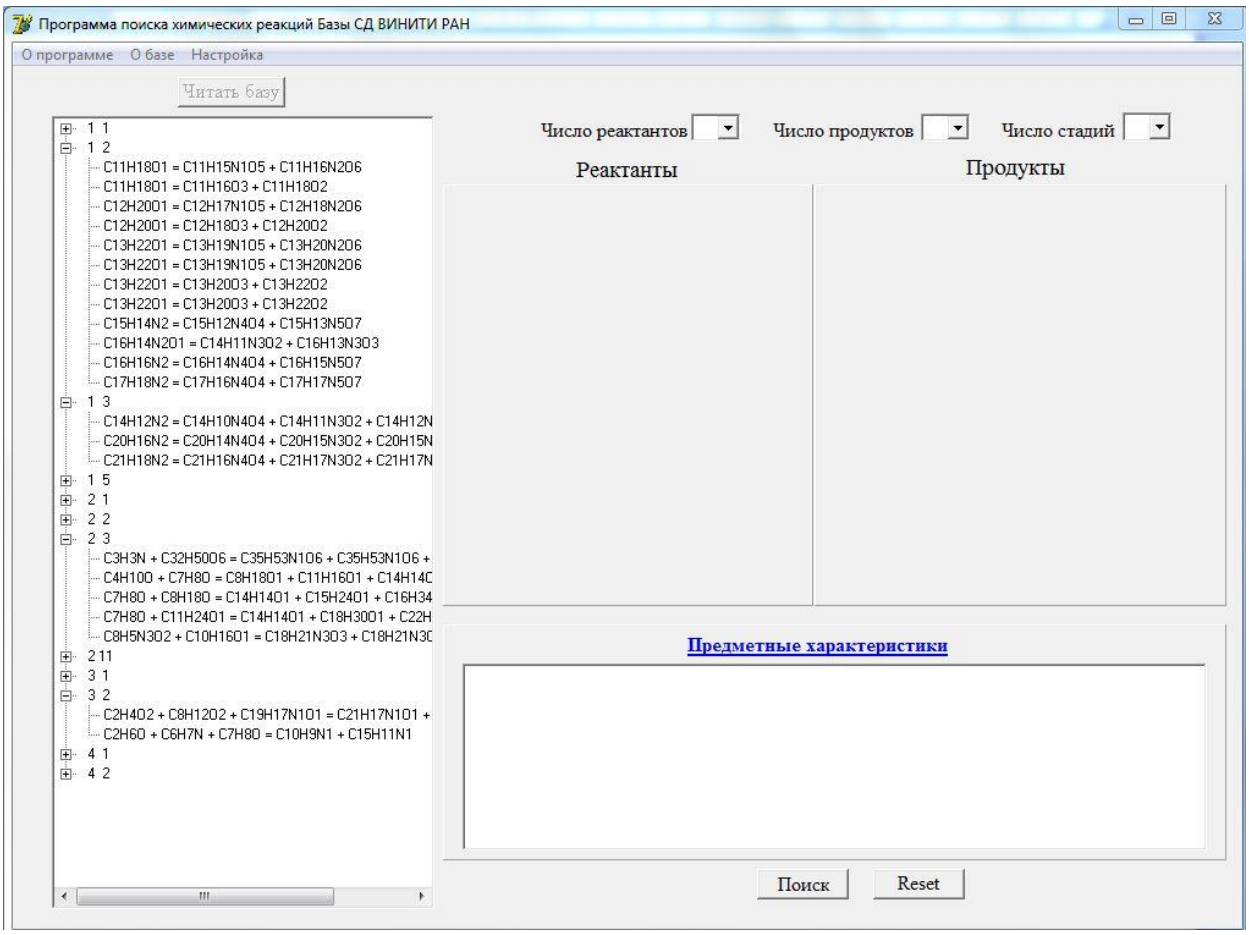

Рис. 2. Интерфейс главного окна программы поиска реакций.

Для поиска в больших пользовательских базах данных предпочтительно находить интересующую информацию с помощью запросов.

Описание видов запросов

Правая часть окна главной формы интерфейса программы поиска реакций (см. рис. 2) предназначена для формирования запросов. Она включает следующие элементы.

1. Выпадающие списки «Число реактантов», «Число продуктов» и «Число стадий». Задание первых двух параметров определяет количества полей в запросах по молекулярным формулам и названиям реактантов и продуктов. Третий параметр позволяет выбрать из базы реакции с заданным числом стадий.

2. Редактируемые поля «Формулы (фрагменты формул)» и «Названия (фрагменты названий)». Сюда можно вводить при помощи клавиатуры или вставлять копированием из текстовых документов соответствующие текстовые строки для реактантов и продуктов (рис. 3).

Из рис. 3 видно, что приведенному запросу удовлетворяет единственная реакция из пользовательской базы:

$$C2H4O2 + C8H12O2 + C19H17N1O1 = C21H17N1O1 + C27H25N1O1.$$

Использование множественного задания запросов, касающихся молекулярных формул и названий реактантов и продуктов усложняет алгоритм поиска, но зато ускоряет получение результатов за счет исключения повторных уточняющих запросов.

3. Поле «Предметные характеристики». Предметная информация о реакциях в Базе СД представлена в форме двух-, трех- или четырехбуквенных предметных характеристик на латинице, построенных по иерархическому принципу [18], каждой из которых соответствует определенное текстовое содержание на русском языке.

При описании реакций используются три группы предметных характеристик:

- группы E – класс реакции (именные реакции, многостадийный синтез, реакции в «одном сосуде» и др.);
- группы Y – различные аспекты изучения и применения реакции (теоретические, физико-химические и технологические; специальные методы синтеза; области применения реакций);
- группы Z – дополнительные условия проведения реакций.

Для того чтобы при запросе получить предметные характеристики, необходимо кликнуть по заголовку поля «Предметные характеристики» (см. рис. 2). Откроется окно (рис. 4), где выбирается нужная группа характеристик, отмечаются необходимые пункты из списка и нажатием кнопки «ОК» выбор подтверждается.

Затем в главном окне в поле «Предметные характеристики» отобразится сделанный выбор (рис. 5).

В результатах такого поиска каждая найденная реакция будет иметь хотя бы одну предметную характеристику из заданного списка.

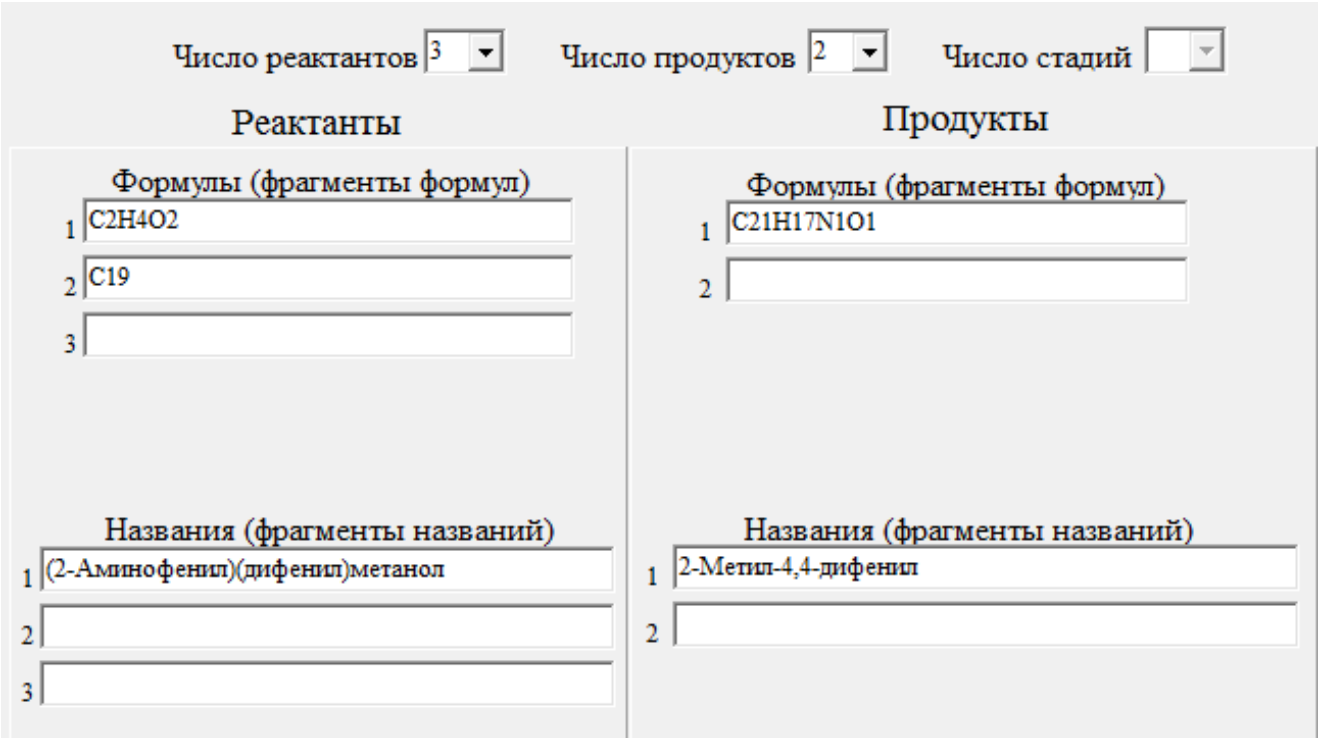

Рис. 3. Пример задания текстовых запросов.

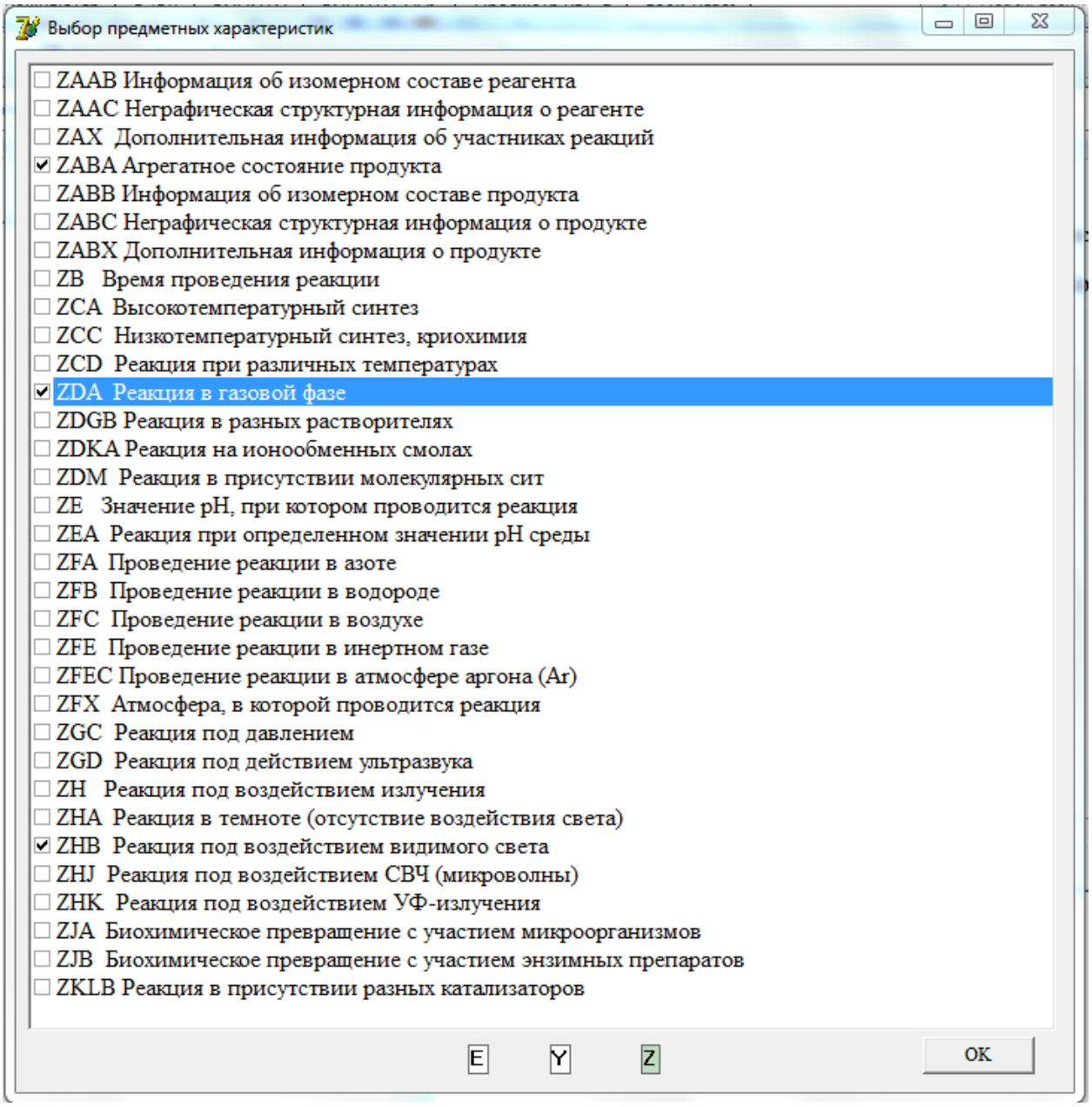

Рис. 4. Окно задания предметных характеристик.

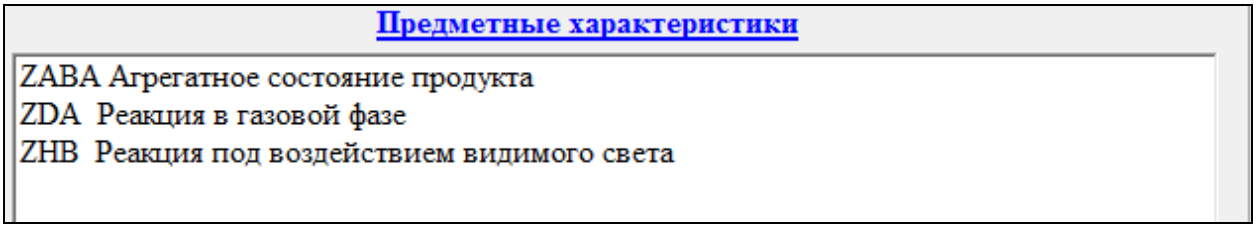

Рис. 5. Пример запроса по предметным характеристикам.

Когда все запросы сформированы, следует нажать кнопку «Поиск», после чего в левой части главного окна отобразится дерево, узлы которого будут соответствовать данным только о найденных реакциях. Если поле, содержащее дерево, окажется пустым, это означает, что по данному запросу ничего не найдено.

Отображение результатов поиска

Просмотр найденной информации о химических реакциях выполняется при помощи дерева реакций. Кликая по узлам дерева с изображением «+» (см. рис. 2), можно раскрывать следующие уровни, содержащие группы записей уравнений. Из уравнений можно по-

пасть в окно «Сведения о реакции» с рисунком уравнения в структурной форме (рис. 6).

В списке продуктов уравнения перед названием продукта реакции указывается его выход в процентах.

В случае наличия больших молекул участников реакции, отображаемых в графическом элементе окна «Сведения о реакции», изображения некоторых атомов и связей могут накладываться друг на друга по причине ограниченности данного графического элемента (рис. 7).

Если пользователю необходимо детально ознакомиться со структурой продукта, он должен кликнуть по названию соединения в списке «Продукты» – откроется окно с соответствующим изображением в увеличенном виде (рис. 8). Таким же способом можно получать изображения структур реактантов из окна «Дополнительно по стадиям».

Окно, приведенное на рис. 9, открывается кликом по строке «Дополнительно по стадиям». В нем приводятся данные по каждой стадии реакции: условия протекания (температура, время, давление), списки реактантов, растворителей, катализаторов и прочих участников.

Для просмотра изображения структуры участника реакции (рис. 10) следует кликнуть по строке с его названием в соответствующем списке.

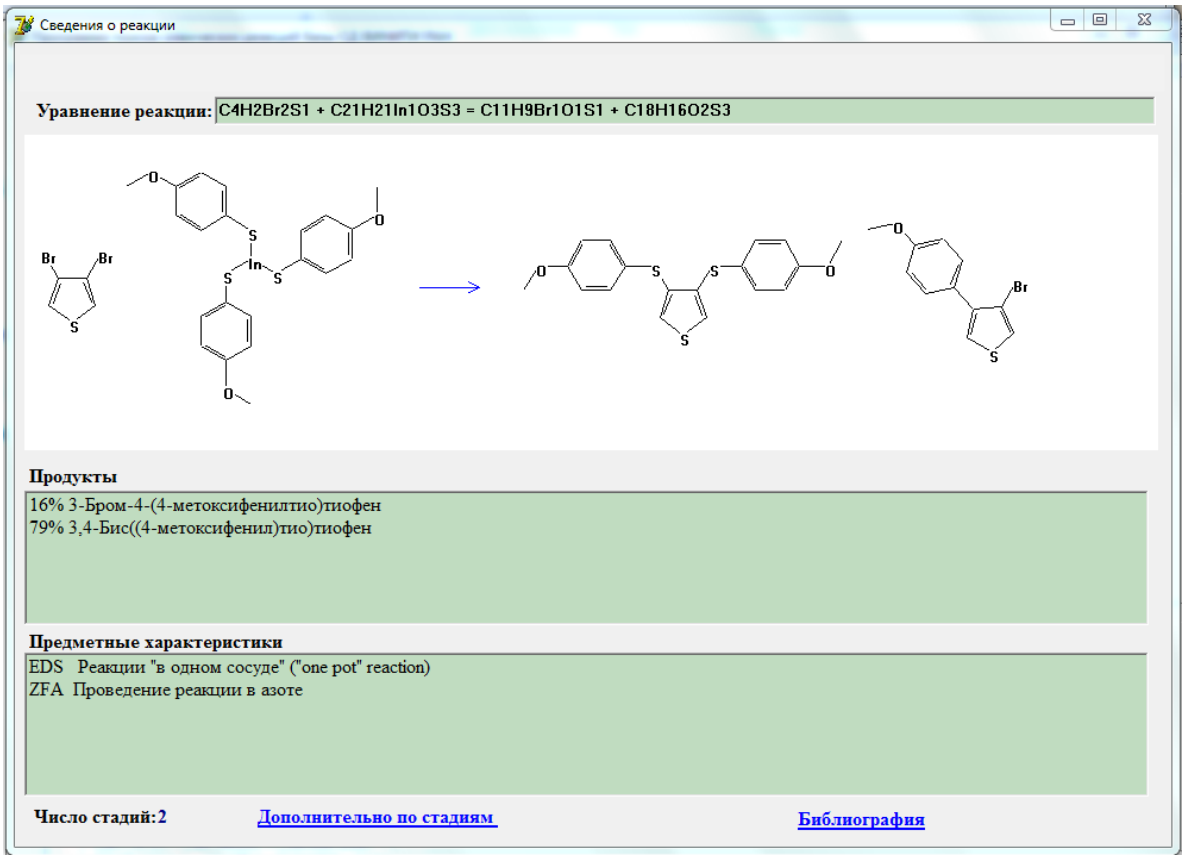

Рис. 6. Окно с общими сведениями о реакции.

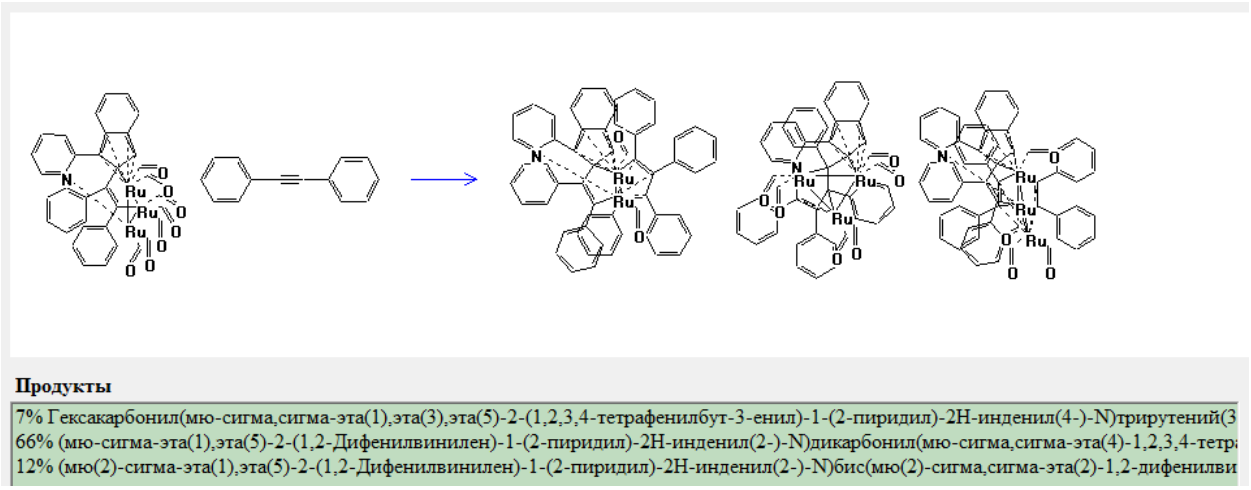

Рис. 7. Пример отображения больших структур в уравнении реакции в графическом виде.

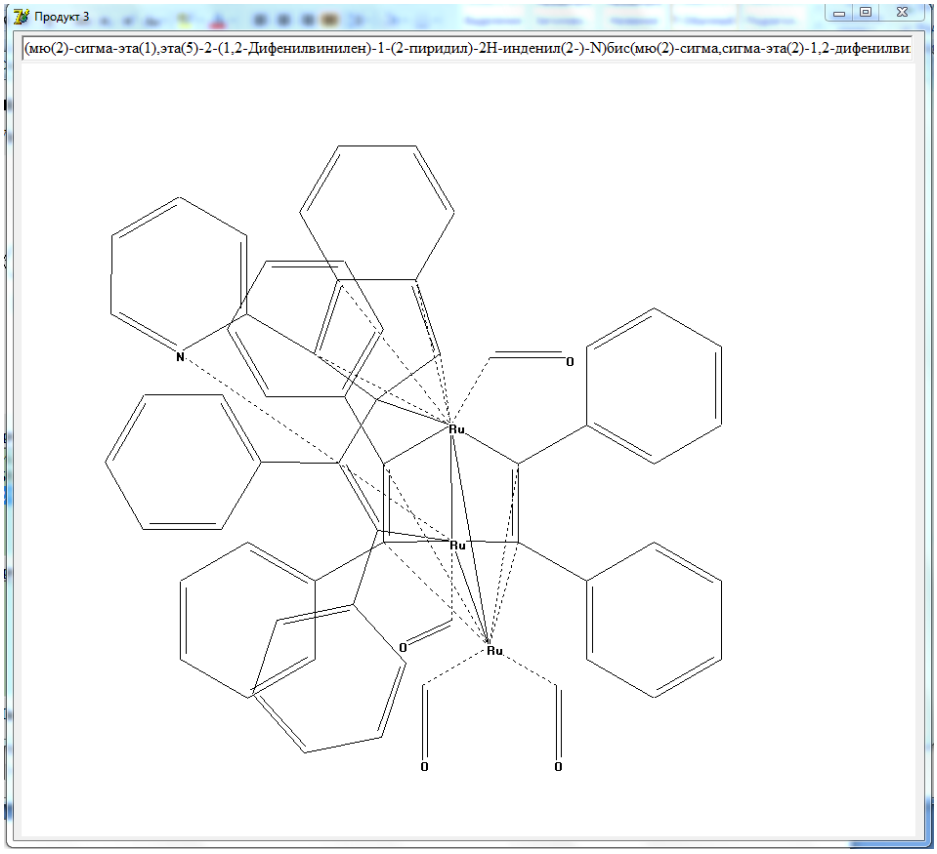

Рис. 8. Увеличенное изображение одного из продуктов реакции (см. рис. 7).

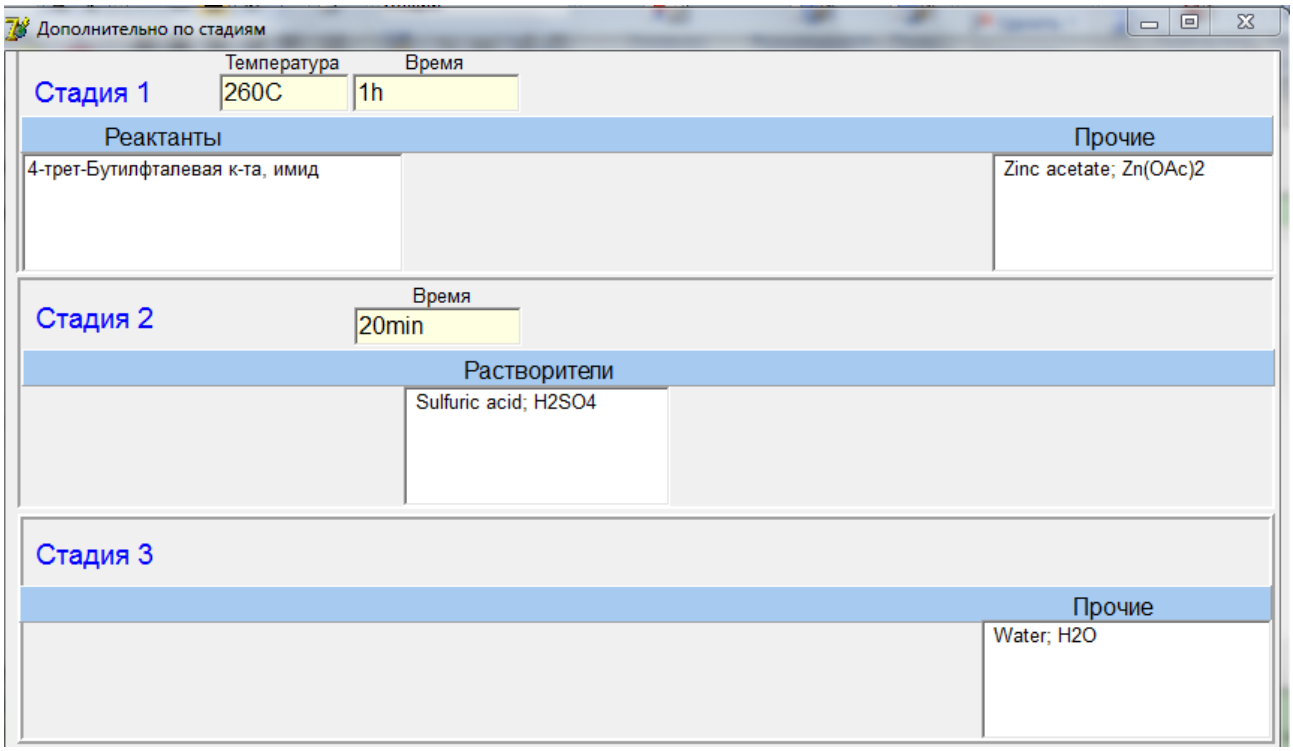

Рис. 9. Пример окна «Дополнительно по стадиям».

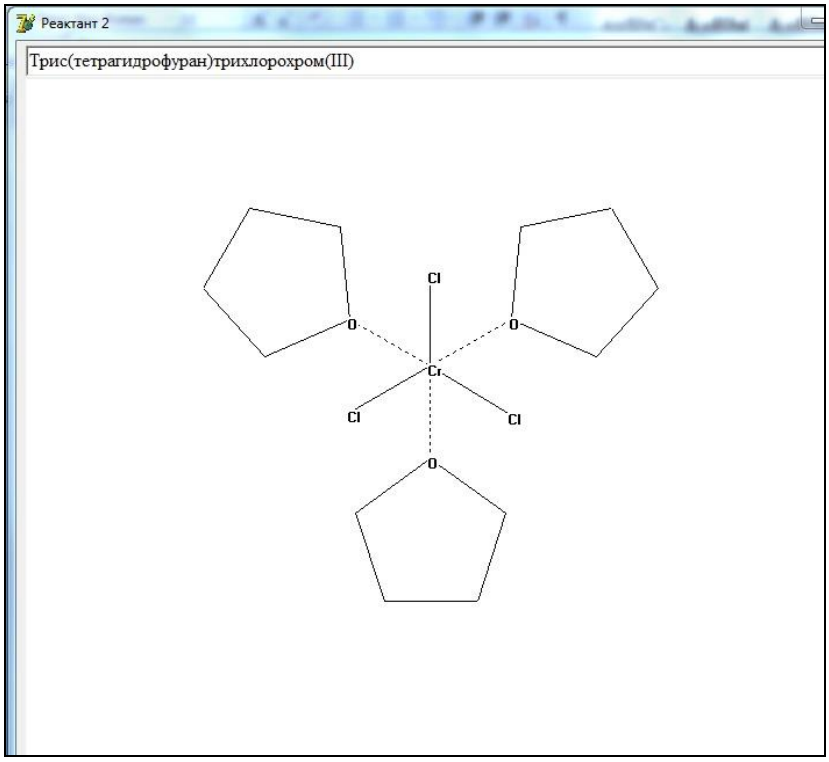

Рис. 10. Структура реактанта.

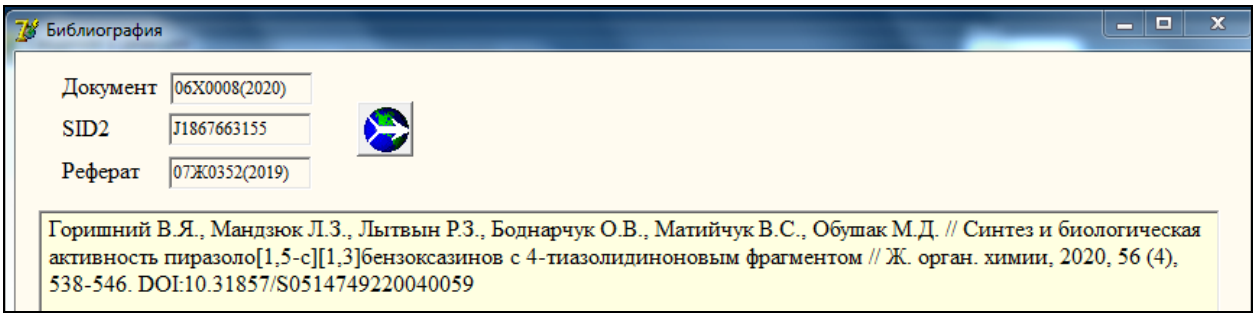

Рис. 11. Сведения о библиографии.

В окне «Сведения о реакции» (см. рис. 6) справа внизу имеется строка Библиография. Щелчок на ней позволяет получить данные об источнике, из которого была получена информация о реакции (рис. 11).

В окне «Библиография» указываются номер документа, идентификатор источника SID2 (см. табл. 1), номер реферата в РЖ «Химия» и библиографическая ссылка. Кнопка справа от поля SID2 предназначена для перехода в Электронный каталог ВИНИТИ РАН с запросом по SID2 для получения подробной библиографической информации по этой публикации. На рис. 12 продемонстрирован результат поиска в Электронном каталоге.

Электронный каталог предоставляет сервис для углубленного библиографического поиска: например, кликнув по фамилии автора, можно получить данные обо всех его публикациях, зарегистрированных в ВИНИТИ РАН.

Об алгоритме выполнения запросов

После старта программы при нажатии на кнопку «Читать базу» упомянутый в подразделе о Конверторе данных бинарный файл General.sdr полностью загружается в оперативную память. При этом формируются элементы компонента TTreeView[4] и отображается дерево реакций базы. Файл General.sdr содержит массив из N записей типа **record**, где N – число реакций в базе. Несложный подсчет позволяет определить верхнюю границу объема оперативной памяти для одной такой записи – 2536 байт. Нами был проведен анализ годовых объемов массивов записей, в результате которого установлено, что максимум числа записей N за один год равен 105615, и получается, что объем необходимой оперативной памяти не превосходит 270 Мб. Этого достаточно, чтобы программа могла работать на обычном персональном компьютере.

[4] Программа написана на Delphi 7.

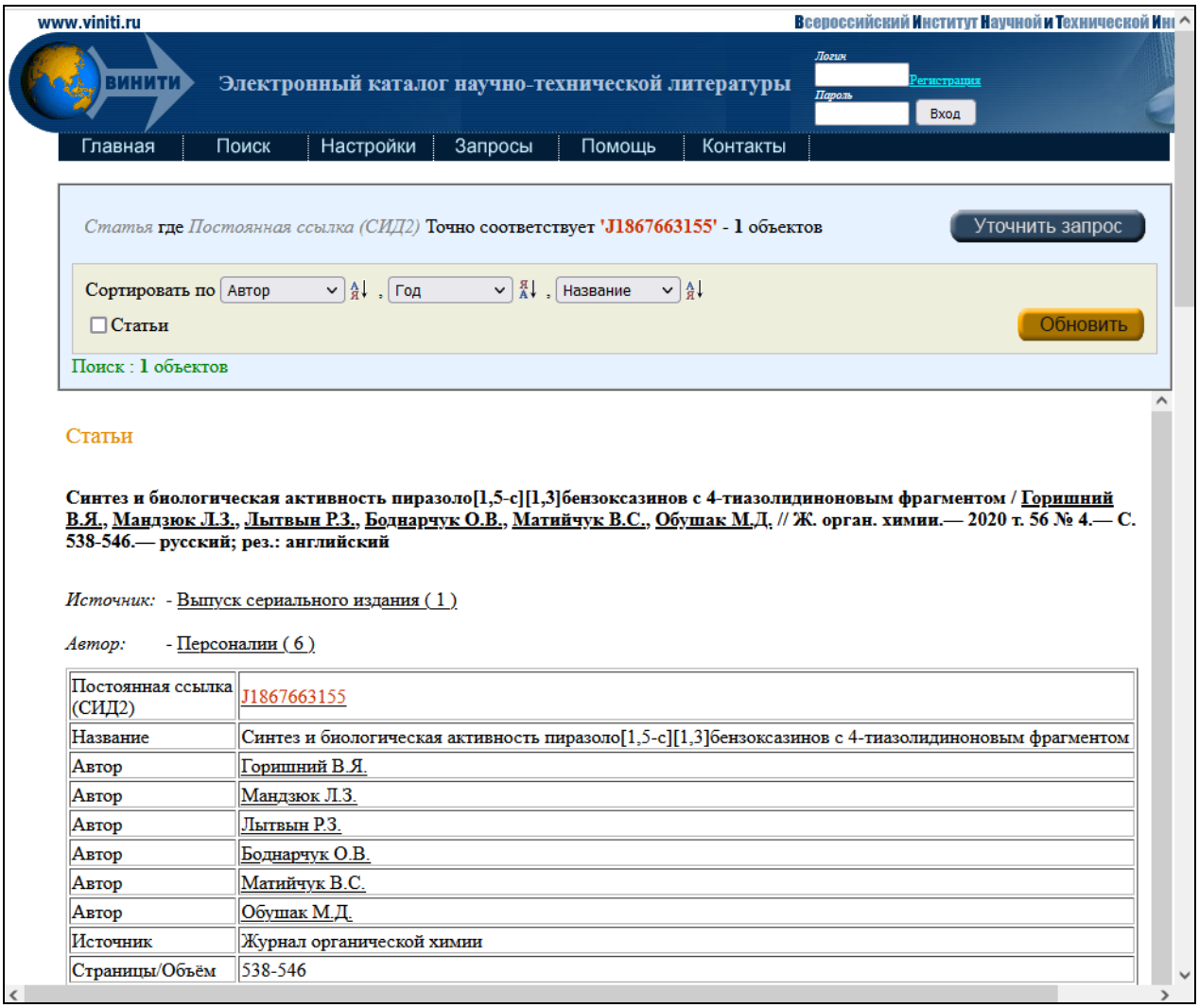

Рис. 12. Пример результата поиска в Электронном каталоге научно-технической литературы ВИНИТИ РАН.

После нажатия на кнопку «Поиск» выполняется просмотр всех записей загруженного в оперативную память файла General.sdr на предмет соответствия заданным запросам. Удовлетворяющие запросам записи отображаются в виде дерева реакций, в котором группы строк с записями уравнений реакций 2-го уровня дерева лексикографически упорядочены. При этом дополнительные операции сортировки не нужны, поскольку все они осуществляются еще на этапе конвертации. Поэтому трудоемкость выполнения поиска по запросу имеет порядок O(N) операций.

Фактор ограниченности объема пользовательской базы данных также имеет большое значение. Конечно, для системы поиска в базах реакций с числом записей порядка 10^6 уже потребуется использование СУБД, оснащенной средствами организации эффективного поиска (индексирование[5], применение уникальных ключей, хеш-таблиц и т.д.), однако этому должны быть посвящены отдельные разработки.

Тестирование системы на массивах данных, содержащих десятки тысяч химических реакций, показало высокую скорость, точность и полноту выполнения запросов. Интерфейс системы удобен и позволяет быстро и наглядно представлять результаты поиска.

Демо-версию автономной системы поиска реакций можно свободно скачать по ссылке https://yadi.sk/d/tH0NMdCRh7-V2w .

ЗАКЛЮЧЕНИЕ

Представленная система поиска информации о химических реакциях – это экспериментальная разработка, способная осуществлять эффективный поиск и наглядно отображать результаты за счет структурной организации пользовательской базы данных. Сегодня она успешно используется в технологических процессах ВИНИТИ РАН и может быть полезна внешним потребителям:

- при проведении фундаментальных и прикладных исследований по химии;
- в учебном процессе;
- в качестве информационного обеспечения баз данных, научных библиотек, издательств.

[5] В рассматриваемом случае индексирование – это использование структур данных, которые помогают СУБД быстрее обнаружить отдельные записи в файле и сократить время выполнения запросов пользователей.

Дальнейшее развитие системы возможно по следующим направлениям.

1. Совершенствование функционала:

- добавление возможности структурного и подструктурного поиска, а также поиска по признакам подобия[6] структур молекул реактантов и продуктов;
- использование в запросах данных о реагентах и прочих участниках реакций;
- создание дополнительных условий поиска, включающих указание на структурные особенности участников реакций (стереохимических особенностей, наличие изотопов, зарядов, радикалов и т.д.);
- поиск по библиографическим данным.

2. Совершенствование структуры пользовательской базы данных за счет:

- создание новых атрибутов химических реакций, реактантов и продуктов с целью структурного поиска (например, InChiKey[7]);
- разработка новых уровней в дереве химических реакций с целью облегчения навигации по данным базы.
- Создание тематических баз данных, например:
- по множествам реакций, определяемым наборами реакционных предметных характеристик;
- по свойствам участников реакций, определяемым предметными характеристиками соответствующих химических структур.

3. Создание интерактивных версий системы поиска химических реакций.

4. Увеличение объемов пользовательской базы данных за счет объединения годовых массивов записей.

СПИСОК ЛИТЕРАТУРЫ

1. Королева Л.М., Федоровская М.А., Чуракова Н.И. и др. Индексирование и ввод сведений о химических соединениях при подготовке базы структурных данных по химии с использованием программного комплекса CBASE32. Инструкция ВИНИТИ РАН 81-2010. – Москва: ВИНИТИ РАН, 2010. – 103 с.
2. Воронежева Н.И., Чуракова Н.И., Нечаева К.С., Пудова Т.А., Немировская И.Б., Трепалин С.В. Индексирование и ввод химических реакций с помощью программы графической обработки данных CBASE. Временная инструкция ВИ 21-97. – Москва, 1997 – 89 с.
3. Воронежева Н.И., Трепалин С.В., Чуракова Н.И., Нечаева К.С., Королева Л.М. Глоссарий как элемент стандартизации ввода данных и программном комплексе CBASE32 // Научно-техническая информация. Сер.2. – 2007. – № 6. – с. 19-24; Voronezheva N.I., Trepalin S.V., Churakova N.I., Nechayeva K S, Koroleva L.M. Glossary as an Element of Data Input Standardization in the Cbase32 Program Complex // Automatic Documentation and Mathematical Linguistics. – 2007. – Vol. 41, № 3 – P. 124-129.
4. Королева Л.М., Н.И.Чуракова, Федоровская М.А., Бессонов Ю.Е., Кирьянова Н.С., Фельдман Б.С., Трепалин С.В. Использование АРМ «Администратор глоссария» при актуализации базы структурных данных ВИНИТИ РАН. – Москва: ВИНИТИ РАН. 2014. – 15 с., ил. – Деп. в ВИНИТИ 14.04.2014, №95-В2014.
5. Именные реакции. – URL: http://orgchemlab.com/name-reactions.html.
6. Ли Дж.Дж. Именные реакции. Механизмы органических реакций. – Москва: БИНОМ. Лаборатория знаний, 2006. – 456 с.
7. Вацуро К.В., Мищенко Г. Л. Именные реакции в органической химии. — Москва: Химия, 1976. – 528 с.
8. Нефедов О.М., Трепалин С.В., Королева Л.М., Бессонов Ю.Е.. Быстрый поиск точных химических структур в больших базах данных с использованием InChI Key кодировки структур // Научно-техническая информация. Сер. 2. – 2013. – № 12. – С. 27-33.
9. Нефедов О.М., Трепалин С.В., Королева Л.М., Бессонов Ю.Е., Чуракова Н.И. База структурных данных по химии ВИНИТИ РАН: проблемы поиска по фрагменту структуры // Научно-техническая информация. Сер. 2. – 2014. – № 12. – С. 19-29.
10. Трепалин С.В., Бессонов Ю.Е., Фельдман Б.С., Кочетова Е.В., Чуракова Н.И., Королева Л.М. База структурных данных по химии ВИНИТИ РАН. Автономная система структурного поиска // Научно-техническая информация. Сер. 2. – 2018. – № 11. – С. 23-31; Trepalin S.V., Bessonov Yu.E., Fel'dman B.S., Kochetova E.V., Churakova N.I., Koroleva L.M. The Structural Chemical Database of the All-Russian Institute for Scientific and Technical Information, Russian Academy of Sciences. An Autonomous System for Structural Searches // Automatic Documentation and Mathematical Linguistics. – 2018. – Vol. 52, № 6. – 297-305.
11. Бессонов Ю.Е., Трепалин С.В., Фельдман Б.С., Кочетова Е.В., Чуракова Н.И. База структурных данных по химии ВИНИТИ РАН. Программы для визуального отображения информации о химических реакциях // Тезисы доклада на XXV Международной научно-методической конференции «Современное образование: содержание, технологии, качество» (г. Санкт-Петербург, 23 апреля 2019 г.). – Санкт-Петербург: Изд-во СПбГЭТУ «ЛЭТИ», 2019. – С. 278-279.
12. Описание MDL. – URL: https://en.wikipedia.org/wiki/MDL_Information_Systems
13. Бессонов Ю.Е., Фельдман Б.С., Чуракова Н.И., Кочетова Е.В., Кирьянова Н.С., Плеханова Н.С. Программы для визуального отображения информации о химических соединениях и реакциях Базы структурных данных по химии ВИНИТИ РАН. – Москва: ВИНИТИ РАН, 2021. – 19 с., ил. Деп. в ВИНИТИ 14.09.2021, №54-В2021.
14. Вход в базу данных Reaxis. – URL: https://www.reaxys.com/

[6] Структурно подобное соединение может содержать, например, позиционные изомеры, другие заместители, и циклические системы. В англоязычной литературе для обозначения структурно подобных веществ используется термин similarity.

[7] InChiKey – международный химический идентификатор, в настоящее время для него отведено поле в Базе Данных ВИНИТИ РАН.

15. Сайт CAS с доступом в систему SciFinder. – URL: https://sso.cas.org/as/cI1zn/resume/as/authorization.ping
16. Инструкция пользователя системы SciFinder. – URL: http://catalysis.ru/resources/info_acenter/Scifinder.pdf
17. Зибарева И.В. Поиск химической информации в базах данных сети STN International. Учебно-методическое пособие для студентов Китайско-российского института. – Новосибирск: Национальный исследовательский университет – Новосибирский государственный университет, Институт катализа им. Г.К. Борескова Сибирского отделения Российской академии наук, 2015. – 81 с.
18. Воронежева Н.И., Трепалин С.В., Чуракова Н.И., Нечаева К.С., Королева Л.М. Система представления и ввода информации о многостадийных химических реакциях с помощью программного комплекса CBASE32 // Научно-техническая информация. Сер. 2. – 2005. – № 7. – С. 7–11.

Материал поступил в редакцию 12.01.22.

Сведения об авторах

БЕССОНОВ Юрий Ефимович – кандидат технических наук, ведущий научный сотрудник ВИНИТИ РАН, Москва
e-mail: bessonov-ye@rambler.ru

ФЕЛЬДМАН Борис Семенович – старший научный сотрудник ВИНИТИ РАН
e-mail: bsf@inbox.ru

ЧУРАКОВА Наталия Исааковна – кандидат химических наук, зав. Отделом исследований и обработки структурной химической информации ВИНИТИ РАН
e-mail: nichurak@rambler.ru

КОЧЕТОВА Елена Вениаминовна – кандидат химических наук, старший научный сотрудник ВИНИТИ РАН
e-mail: kelena63@gmail.com