УДК 004.35:81'374.3

Н. И. Воронежева, С. В. Трепалин, Н. И. Чуракова,
К. С. Нечаева, Л. М. Королева

Глоссарий как элемент стандартизации ввода данных в программном комплексе CBASE32*

Разработан Глоссарий — справочник химических соединений, являющийся компонентом 32-х разрядной версии программы графической обработки структурных данных CBASE32. Глоссарий позволяет стандартизировать представление структурных и фактографических данных о химических соединениях и упростить процедуру ввода соединений и реакций в Базу структурных данных по химии ВИНИТИ.

ВВЕДЕНИЕ

Создание структурных баз по химии является сложной и кропотливой работой, требующей от химика разносторонних знаний и специальной подготовки. В процессе работы химики вынуждены обращаться к большому числу справочных изданий.

Современное развитие информационных технологий позволяет при создании баз структурных данных по химии использовать электронные словари-справочники (электронные глоссарии). Наиболее удачными примерами таких справочников являются электронные издания Chemical Abstracts, а также электронный аналог хорошо известного каталога фирмы Merck, позволяющие работать как с текстовыми данными, так и с химическими структурами. Однако эти электронные издания являются малодоступными и дорогостоящими, и не позволяют обеспечивать создателей баз данных структур химических соединений, в частности, создателей Базы структурных данных по химии ВИНИТИ, специфичной русскоязычной информацией. Кроме того, в перечисленных электронных изданиях не предусмотрено расширение набора данных, а также внесение каких-либо изменений и дополнений к имеющимся сведениям. Это вынуждает химиков формировать свои собственные электронные каталоги химических соединений (структур, номенклатурных названий и свойств), отвечающие требованиям, предъявляемым к базе данных.

В статье описан электронный химический справочник — Глоссарий, который предназначен для создания двух составляющих базы структурных данных по химии ВИНИТИ — собственно базы структур химических соединений, и базы реакций, в которых эти соединения принимают участие.

В работе [1] сообщалось о первой отечественной графической программной оболочке CBASE (16-разрядная версия), предназначенной для ввода и обработки структурных, фактографических и библиографических данных по химии. Программная оболочка CBASE (Chemical Base) позволяет осуществлять ввод, хранение и поиск информации о химических соединениях и реакциях, в которых эти соединения принимают участие [2–5]. Программный комплекс CBASE32 является развитием программной оболочки CBASE и представляет ее 32-разрядную версию [6]. Новым существенным компонентом этого программного комплекса является справочник химических соединений — Глоссарий, которому и посвящена настоящая статья.

СТРУКТУРА И НАЗНАЧЕНИЕ ГЛОССАРИЯ

В созданном нами Глоссарии объединены и структурированы сведения о химических соединениях из различных источников: База структурных данных по химии ВИНИТИ [1], Справочник химика [7], справочные издания CAS [8].

Глоссарий содержит разнообразные сведения о химических соединениях, к которым относятся:

— систематическое название (SN),
— синонимическое название (SYNONYMS),
— аббревиатура (ABBR),
— молекулярная формула (BF),
— идентификационный номер в Глоссарии (II),
— регистрационный номер CAS,
— регистрационный номер соединения в Базе структурных данных по химии ВИНИТИ (VINITI),
— структура соединений,
— комментарий (COMMENT).

Систематическое название дается по Номенклатуре IUPAC. **Синонимическое название** может включать тривиальное название и несколько синонимов на русском и английском языках. В поле **<ABBR>** занесены русскоязычный и англоязычный варианты сокращений названия химического соединения и его линейная формула. **Линейная формула** приводится в соответствии с правилами представления реагентов в программе графической обработки структурных данных CBASE [3]. Порядок следования химических элементов в **молекулярной формуле** соответствует порядку букв в латинском алфавите (для соединений, не содержащих атомов углерода). В молекулярной формуле соединений с углеродом на первом месте стоит

*Работа поддержана грантом № 04-07-90126b Российского фонда фундаментальных исследований.

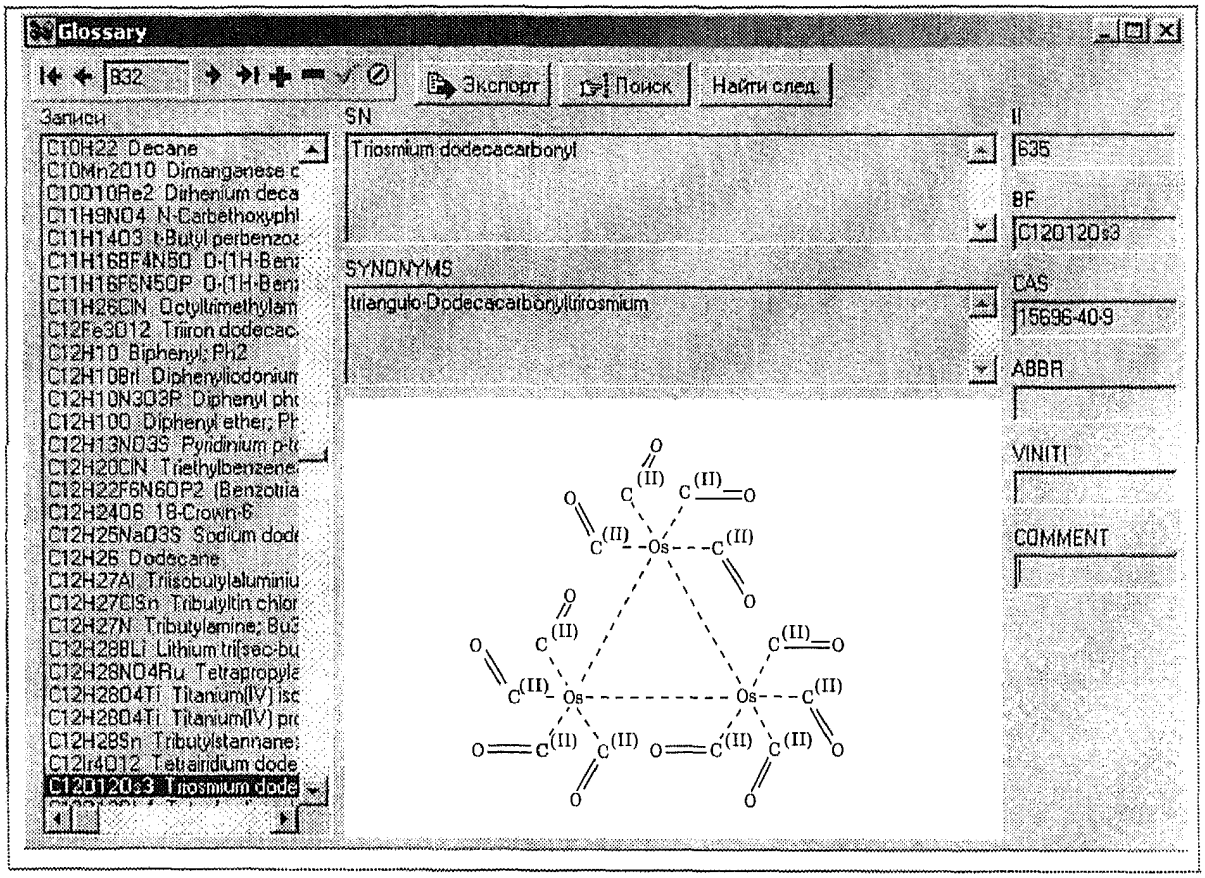

Рис. 1. Вид окна Глоссария для администратора

углерод, затем водород (если он есть), затем все остальные элементы в алфавитном порядке. По такому же принципу упорядочены молекулярные формулы соединений из Глоссария при сортировке их по молекулярной формуле. Содержание полей **<регистрационный номер CAS>**, **<регистрационный номер соединения в Базе структурных данных по химии ВИНИТИ>** очевидно и не требует специальных пояснений. **Структура соединения** приводится в соответствии с требованиями к представлению структур [9] в Базе структурных данных по химии ВИНИТИ (База СД).

Как уже было сказано, Глоссарий используется для создания базы структур химических соединений и базы реакций, в которых эти соединения принимают участие. При создании базы структурных данных Глоссарий служит для ввода часто встречающихся соединений, позволяя эксперту-химику при обработке информации о химических соединениях выбирать готовые структуры с названиями. Это значительно облегчает ввод сложных структур и позволяет избежать ошибок при построении структур и написании их названий. Кроме того, в Глоссарии содержатся родоначальные представители некоторых классов соединений, которые могут использоваться в качестве заготовок для ввода замещенных соединений, а также некоторые уникальные, редко встречающиеся соединения, поиск которых в справочной литературе затруднителен. Пример такого сложного комплексного соединения приведен на рис. 1. Таким образом, Глоссарий позволяет не только увеличить производительность труда создателей баз данных химических соединений, но и стандартизировать представления химических структур и их названий в создаваемых базах.

Глоссарий используется также при вводе реакций. Из него берутся соединения, которые необходимы для описания уравнения химической реакции, но не являются предметом изучения в научной статье, что позволяет существенно сократить объем работы по созданию базы реакций. Кроме того, соединения из Глоссария являются компонентами списков растворителей, катализаторов и прочих участников реакции. Например, для ввода неиндексируемого участника реакции с аббревиатурой DCC, эксперту-химику достаточно найти это соединение по указанной аббревиатуре, и с его порядковым номером в Глоссарии это соединение будет включено в уравнение реакции.

О ПРОГРАММЕ

Глоссарий является компонентом программы CBASE32, являющейся приложением, которое устанавливается на каждом клиентском месте. Глоссарий представляет собой самостоятельную базу данных. Особенностью этой базы следует считать прямую работу с SD-файлом без его конвертирования в какие-либо двоичные форматы хранения данных. Это замедляет время работы с данными, но делает базу более универсальной, поскольку формат SD-файла описан [10] и де-факто является стандартом обмена структурными данными. Все коммерческие программные продукты,

работающие с химическими структурами, имеют возможность экспорта в SD-файл. Такая возможность предусмотрена и для базы данных Глоссария. Все поля, хранящиеся в SD-файле, детектируются автоматически и могут быть доступны для показа пользователю, для поиска и сортировки.

Программа работы с Глоссарием написана на языке Delphi в виде отдельного модуля и является переносимой, т. е. Глоссарий легко использовать не только в CBASE32, но и в других приложениях, работающих с химическими данными. При создании Глоссария использовались библиотеки низкого уровня [11]. Программа написана для Win32 платформ.

ПРИНЦИП СОЗДАНИЯ И ПОПОЛНЕНИЯ ГЛОССАРИЯ. РАБОТА АДМИНИСТРАТОРА С ГЛОССАРИЕМ

Основу Глоссария составляет список соединений, отобранных по частоте встречаемости и значимости из накопленного массива (более трех миллионов соединений) Базы СД за 10 лет.

Для пополнения Глоссария в программном комплексе CBASE32 предусмотрено создание дополнительного списка соединений. Соединения, предназначенные для включения в этот список, отбираются экспертами-химиками в процессе создания Базы СД. Для этой цели предусмотрен уникальный элемент описания — предметный терм. Химик-эксперт присваивает потенциальному компоненту Глоссария этот предметный терм. Затем извлеченные по данному признаку соединения анализируются администратором по частоте встречаемости, сложности и значимости. Структуры и названия выбранных соединений подвергаются тщательной проверке и после двукратного независимого редактирования экспертами-химиками включаются администратором в Глоссарий.

Ввод структуры нового химического соединения в Глоссарий осуществляется с помощью структурного редактора, встроенного в программный комплекс CBASE32. Для удобства работы администратора в Глоссарий встроена поисковая система.

В этой системе возможен поиск по всем элементам описания химического соединения в Глоссарии. Создаваемая администратором база данных химических соединений в Глоссарии может быть экспортирована в стандартном формате SD-файла [10].

Уровень доступа к Глоссарию определяется правами пользователя: администратор может вносить изменения в Глоссарий, пользователь-редактор — только копировать соединения из Глоссария в текущую базу данных. Администратору доступно редактирование всех полей Глоссария, а также добавление новых записей или их удаление в Глоссарии. На рис. 1 приведен вид окна Глоссария для администратора.

РАБОТА ЭКСПЕРТА-ХИМИКА С ГЛОССАРИЕМ

Эксперт-химик, обрабатывая информацию о химических соединениях в программной оболочке CBASE32, создает фрагмент Базы СД. При этом он может копировать соединения из Глоссария в текущую базу данных химических соединений или реакций (рис. 2). При работе с Глоссарием пользователю предоставлена возможность самому выбрать набор полей для визуализации, воспользовавшись опцией <Вид>.

ИСПОЛЬЗОВАНИЕ ГЛОССАРИЯ ПРИ ВВОДЕ СОЕДИНЕНИЙ

Эксперт-химик вызывает Глоссарий в окне СОЕДИНЕНИЕ и находит в Глоссарии нужное ему химическое соединение (рис. 3). При выходе из Глоссария структура соединения и другие его атрибуты автоматически заносятся в соответствующие поля окна текущего соединения.

Для выбора нужного соединения в Глоссарии реализованы следующие виды поиска:

1. Быстрый поиск по началу слова. Выбрав нужную колонку для поиска, пользователь вводит в контроль *Начало слова* необходимые символы. При этом поиск осуществляется немедленно, и, если будет найдена запись, которая в начале строки содержит введенные символы, то курсор в Глоссарии немедленно позиционируется на этой записи. Если не будет найдена запись с данным началом слова, то положение текущей записи не изменится.

2. Поиск по фрагменту текста. После выбора колонки для поиска в контроль *Фрагмент Текста* вводится искомый фрагмент текста. Поиск осуществляется, начиная с текущей записи. Как только будет найдена запись, которая содержит заданный фрагмент текста, курсор будет позиционироваться на ней. При повторном нажатии на кнопку *Поиск*, будет найдена следующая запись. Если будут просмотрены все записи, включая последнюю и фрагмент текста не будет найден, поиск будет продолжен, начиная с первой записи. При отсутствии заданного фрагмента текста пользователь будет предупрежден.

3. Поиск по названию соединения. Осуществляется по фрагменту текста, заданному в контроле *Фрагмент текста*. В отличие от раздела 2, поиск одновременно осуществляется в полях SN, SYNONYMS и ABBR, поэтому не требуется выбирать колонку для поиска. Если фрагмент текста будет найден хотя бы в одном из указанных выше полей, курсор позиционируется на эту запись.

4. Поиск по фрагменту структуры. Осуществляется с помощью структурного редактора, встроенного в программный комплекс CBASE32. Поиск стартует, начиная с текущей записи, и как только будет найдена запись со структурой, содержащей данный фрагмент, курсор будет позиционироваться на ней. Повторное нажатие кнопки *Поиск* найдет следующую запись и так далее. После того как будет найдена последняя запись, поиск продолжится с первой записи.

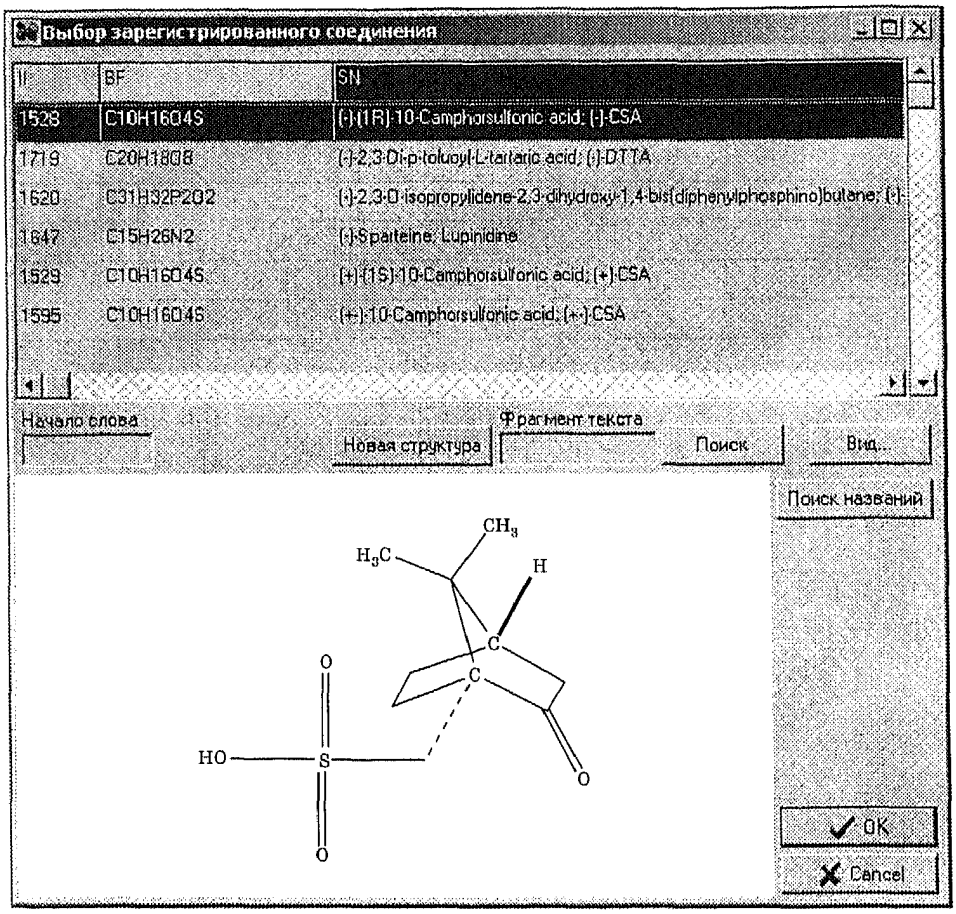

Рис. 2. Вид окна Глоссария для эксперта-химика

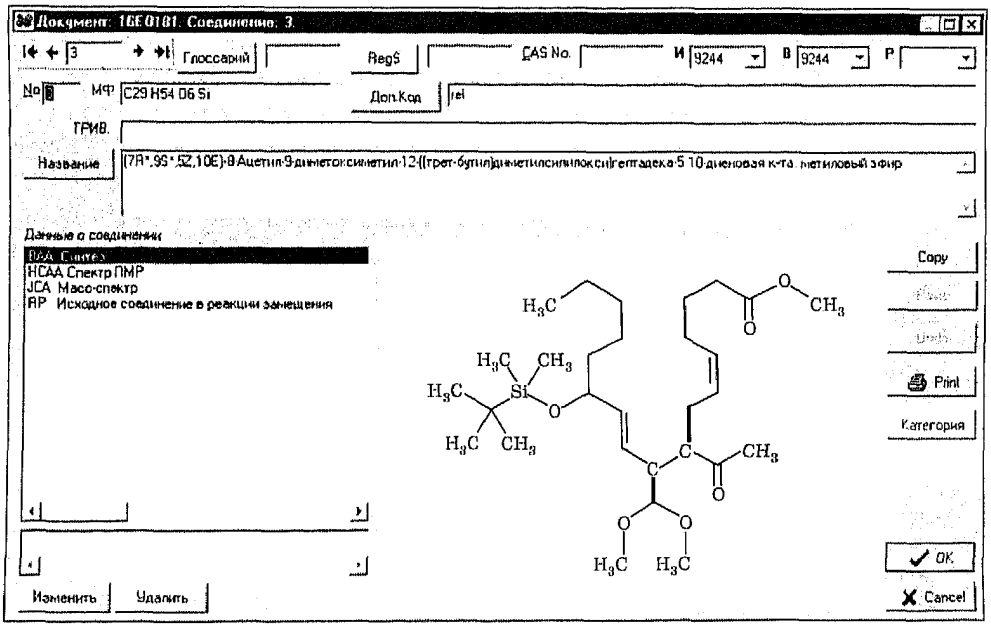

Рис. 3. Вид окна СОЕДИНЕНИЕ с кнопкой вызова Глоссария

ИСПОЛЬЗОВАНИЕ ГЛОССАРИЯ ПРИ ВВОДЕ РЕАКЦИЙ

Согласно принятой в программном комплексе CBASE32 системе представления и ввода информации о реакциях [6], участниками реакции являются реактанты, продукты, растворители, катализаторы и прочие участники реакции (т. е. соединения, не относящиеся к уже названным участникам реакции, но способствующие ее протеканию). Схема реакции представляется в базе реакций в виде уравнения реакции, в левой части которого отражаются номера соединений-реактантов, а в правой номера соединений-продуктов. Остальные участ-

ники реакции в уравнении реакции не отражаются.

При вводе информации о реакциях в базу реакций Глоссарий может использоваться по-разному, в зависимости от той роли, которую играет данное соединение в реакции. Во-первых, он служит для включения в уравнение реакции неиндексируемых участников реакции. Для этого в окне ввода реакции предусмотрена возможность выбора ее компонентов из Глоссария (рис. 4). Каждое выбранное из Глоссария соединение включается в уравнение реакции под его идентификационным номером в Глоссарии. Чтобы отличить номера соединения из Глоссария от номеров, заиндексированных в текущем документе соединений, перед первыми в уравнении реакции ставится звездочка (*). Каждому выбранному из Глоссария реагенту пользователь приписывает номер стадии, на которой данный реагент вводится в реакцию. Номер стадии указывается после косой черты следом за номером соединения из Глоссария. По умолчанию номеру стадии приписывается единица, и в таком случае он не отражается в уравнении реакции.

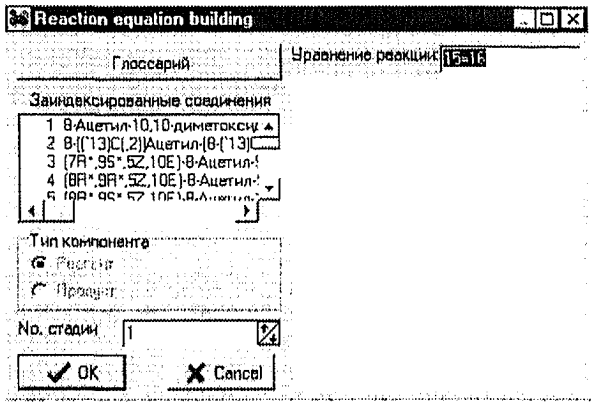

Рис. 4. Вид окна РЕАКЦИЯ с кнопкой вызова Глоссария

Кроме того, Глоссарий используется для ввода в запись реакции неиндексируемых компонентов реакции — растворителей, катализаторов и прочих участников реакции. При задании такого участника реакции пользователю предлагается обратиться к Глоссарию и выбрать из него нужное соединение. Предлагаемое пользователю в данном случае обращение к Глоссарию аналогично обращению к нему в окне СОЕДИНЕНИЕ, с той лишь разницей, что каждому выбранному растворителю, катализатору или прочему участнику реакции приписывается номер стадии, на которой он вводится в реакцию.

РАБОЧЕЕ МЕСТО ПОЛЬЗОВАТЕЛЯ

Теоретически программный комплекс CBASE32, включающий встроенную базу данных Глоссария, может работать на любом компьютере с операционной системой Windows (95, 98 или NT). Однако для достижения удовлетворительных результатов компьютер должен иметь:

— Processor Pentium 500 или лучше;

— 32Mb RAM для запуска самого CBASE32. Отметим, что CBASE32 создает собственную временную индексацию в памяти для открытой базы. Максимальный размер индекса — 24 байта на структуру;

— 5Mb на жестком диске для установки. Дополнительная память зависит от размера используемых баз данных. В среднем, одна химическая структура занимает 512 байт на диске;

— "мышь", совместимую с операционной системой Windows.

ЗАКЛЮЧЕНИЕ

В настоящее время в Глоссарий занесено более 1700 химических соединений. В процессе эксплуатации программного комплекса CBASE32 происходит постоянное пополнение Глоссария. Однако следует иметь в виду, что значительное увеличение количества соединений в Глоссарии приведет к увеличению времени обращения к нему и необходимого для программного комплекса CBASE32 свободного пространства на компьютере, а также потребует увеличения быстродействия используемого компьютера.

СПИСОК ЛИТЕРАТУРЫ

1. Алфимов М. В., Авакян В. Г., Трепалин С. В., Воронежева Н. И., Чуракова Н. И. Универсальная программная оболочка для создания баз данных химических соединений и реакций // Доклады РАН.— 1999.— Т. 366, № 5.— С. 639–642.

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 структурных данных CBASE // Временная инструкция: ВИ 21-97.— М.: ВИНИТИ, 1997.— 89 с.

3. Трепалин С. В., Авакян В. Г., Воронежева Н. И., Чуракова Н. И., Нечаева К. С., Пудова Т. А., Немировская И. Б. Представление химических реакций в программе графической обработки структурной химической информации CBASE // Материалы 3-й Междунар. конф. "НТИ-97: Информационные ресурсы. Интеграция. Технологии".— М., 1997.— С. 201–202.

4. Трепалин С. В., Авакян В. Г., Воронежева Н. И., Чуракова Н. И. *Поиск реакций по изменяющимся фрагментам* в программе графической обработки структурной химической информации CBASE // Материалы 3-й Междунар. конф. "НТИ-97: Информационные ресурсы. Интеграция. Технологии".— М., 1997.— С. 9–12.

5. Воронежева Н. И., Чуракова Н. И., Нечаева К. С., Пудова Т. А., Немировская И. Б., Трепалин С. В., Тертерян Р. А. Основные принципы организации и наполнения базы реакций в Банке структурных данных ВИНИТИ по химии // Материалы 4-й Междунар. конф. "НТИ-99: Интеграция. Информационные технологии. Телекоммуникации".— М., 1999.— С. 72–77.

6. 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 // НТИ. Сер. 2.— 2005.— № 7.— С. 7–11.