Всероссийский институт научной и технической информации РАН (ВИНИТИ РАН), г. Москва, Россия

Аннотация. *Рассмотрен один из аспектов информационной деятельности ВИНИТИ РАН, связанный с исследованиями в области химии и химической технологии. Дано описание методов реализации двух направлений по данной тематике – формирования хранилища химической информации и предоставления накопленной информации потребителям.*

Ключевые слова: химическая информация; базы структурных данных по химии; структурный поиск в режимах online и offline

В формировании глобального информационного общества большую роль играют научные информационные институты. ВИНИТИ РАН – это крупнейший научно-информационный и аналитический центр России, который обеспечивает с 1952 г. российское и мировое сообщество научно-технической информацией по проблемам точных, естественных и технических наук.

Настоящий доклад посвящён одному из аспектов информационной деятельности ВИНИТИ РАН, связанному с исследованиями в области химии и химической технологии.

Информационная деятельность по этой тематике осуществляется в двух направлениях:

– формирование хранилища химической информации;

– предоставление накопленной информации потребителям.

Работа по накоплению информации о химических соединениях и реакциях из потока отечественной и зарубежной научной литературы ведется в ВИНИТИ РАН с 1974г. По состоянию на 2024 г. База структурных данных по химии ВИНИТИ РАН (База СД) содержит информацию о более чем 5,7 млн. химических структур, около 1,7 млн. химических реакций и 15 млн. свойств химических соединений [1]. Ежегодно она пополняется информацией о более чем 30 тыс. соединений и 20 тыс. реакций.

В Базе СД структурной информации о химических веществах и реакциях сопутствуют комментарии, которые являются дополнительными данными об особенностях химических молекул и процессов.

База СД даёт возможность ученому-химику оперативно получать нужную информацию о составе и характеристиках веществ, способах их синтеза, реакциях с их участием, сферах использования (фармакология, защита окружающей среды и др.).

Одним из основных элементов технологического процесса формирования Базы СД является комплектование входного потока документов для аналитической обработки. Спецификой Базы СД является ее направленость на аккумуляцию данных, отраженных в научно-технической литературе по органической химии, и главным образом касающихся проблем органического синтеза. Современное общество нуждается в постоянном создании новых жизненно важных химических веществ с заранее заданными свойствами.

Для формирования информационного хранилища был разработан полностью автоматизированный технологический процесс, который включает три основных этапа:

1. Формирование интегрированного электронного ресурса документов на основе входного потока.

2. Извлечение структурной химической информации из документов, ввод её в Базу СД и научное редактирование введённой информации. Эта работа выполняется квалифицированными специалистами-химиками.

3. Автоматическая проверка корректности введённой в Базу СД информации.

Неоспоримым достоинством данной технологии является исключение работы с использованием бумажных носителей информации на каждом из указанных этапов.

С самого начала формирования Базы СД её содержание было востребовано зарубежными и отечественными потребителями. Данные передавались в специальных обменных форматах.

Несомненно, что в накопленном большом массиве востребованной информации, необходимо обеспечить эффективный поиск, под которым понимается точность, полнота и наглядность результатов при быстром выполнении запросов. Следует отметить, что внутренний формат данных Базы СД предназначен для технологических нужд, а обменные форматы успешно используются для взаимодействия между базами данных, но они, так же как внутренние форматы, не предназначены для ведения эффективного поиска. Поэтому, в 2013г. были начаты разработки эффективных поисковых систем, функционирующих в интерактивном и автономном режимах, основанных на создании рабочих баз данных с иерархическим упорядочением информации о химических структурах и реакциях.

При работе в интерактивном режиме клиент получает через Интернет доступ ко всей базе данных, размер которой может достигать порядка 106 записей. Апробация системы на массиве данных порядка 105 записей показала высокую эффективность поиска [2, 3]. В настоящее время ведутся работы по созданию новых усовершенствованных систем поиска химической информации Базы СД в режиме online.

Альтернативой интерактивного режима доступа является режим offline, при котором клиенту предоставляется пользовательская база данных (ПБД), которая представляет собой локальный массив Базы СД, сформированный то тем или иным признакам (как правило, это календарный период времени формирования массива: месяц, квартал, полугодие, год) [4, 5]. ПБД может быть также сформирована, например, в результате обработки годового объема выпусков одного конкретного научного журнала. Работа автономных систем поиска структур и 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.

ПБД может быть так же тематической – содержащей информацию о химических структурах и реакциях, обладающих заданными характеристиками.

В настоящее время созданы ПБД, содержащие в совокупности более 1,2 млн. записей химических структур и более 0,7 млн. записей химических реакций. Так же сформирована тематическая ПБД комплексных химических соединений на основе данных из научных публикаций 2022-2023гг.

Дополнительную информацию о Базе СД и системам поиска химической информации можно получить на сайте ВИНИТИ РАН по ссылке http://www.viniti.ru/products/bd-sd.

Список литературы:

1. Чуракова Н. И., Бессонов Ю. Е., Фельдман Б. С., Червинская Н. В. База структурных данных по химии ВИНИТИ РАН. Вопросы формирования, эксплуатации и создания информационных продуктов / Н. И. Чуракова, Ю. Е. Бессонов, Б. С. Фельдман, Н. В. Червинская // Научные и технические библиотеки. 2022. № 10. С. 31–51. https://doi.org/10.33186/1027-3689-2022-10-31-51.

2. 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.

3. Нефедов О. М., Трепалин С. В., Королёва Л. М. [и др.] База структурных данных по химии ВИНИТИ РАН: проблемы поиска по фрагменту структуры//Научно-техническая информация. Сер.2. 2014. № 12. С.19–29.

4. Trepalin S. V., Bessonov Yu. E., Fel'dman B. S. [et al.] 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. P. 297–305.

5. Бессонов Ю. Е., Фельдман Б. С., Чуракова Н. И. [и др.] Поиск и отображение информации о химических реакциях в базе структурных данных по химии ВИНИТИ РАН // Научно-техническая информация. Сер. 2. 2022. № 3. С. 10–22.

Yu. E. Bessonov, N. I. Churakova, B. S. Feldman, E. V. Kochetova

Structural data base on chemistry of VINITI RAS. Methods of collecting and disseminating information

Russia Russian Institute for Scientific and Technical Information (VINITI RAS), Russia

Abstract. One of the aspects of the information activities of VINITI RAS related to research in the field of chemistry and chemical technology is considered. A description of methods for implementing two directions on this topic is given – the formation of a repository of chemical information and the provision of accumulated information to consumers.

Keywords: chemical information; structural databases for chemistry; structural search in online and offline modes