# ПОИСК СВЕДЕНИЙ О ХИМИЧЕСКОМ ВЕЩЕСТВЕ В ОНЛАЙНОВЫХ ИНФОРМАЦИОННЫХ ИСТОЧНИКАХ

**Рагойша А.А.**

*Химический факультет Белорусскогой государственного университета*
*Минск, Республика Беларусь*

Поиск сведений о веществе — одна из тех задач, которые химику приходится решать постоянно в своей повседневной деятельности. В текстовых документах, в базах данных одно и то же вещество отображают многочисленными способами, каждый из которых особенно полезен и удобен в своей области использования. В первичной литературе читатель обычно сталкивается с одной-двумя формами идентификаторов (название вещества, химическая формула); в реферативных и справочных базах данных, где накапливается информация из множества источников, перечень идентификаторов вещества может исчисляться десятками. Так например, в Википедии (особенно ее английской версии) на странице вещества мы видим, кроме тривиального и систематического названия и кроме брутто- и структурной формул, еще и большую группу регистрационных номеров, а также коды *SMILES* и *InChI*.

Многообразие форм отображения вещества приходится учитывать при ведении информационного поиска.

Пользователь должен четко понимать, что:

- на данном этапе развития Интернета ни одна из этих форм, примененная в запросе, не обеспечит обнаружение всех имеющихся в наличии релевантных документов;
- каждый тип информационных источников тяготеет к использованию некоего характерного набора идентификаторов;
- у каждого способа идентификации вещества имеются свои достоинства, недостатки, особенности и, следовательно, цели применения.

На химическом факультете Белорусского государственного университета в рамках курса «Информационные технологии в химии» мы знакомим студентов с теми идентификаторами химических веществ, которые часто встречаются в онлайновых источниках и которые могут быть полезны в информационном поиске [1].

Алгоритм обнаружения информации по известному названию вещества в основной своей части не отличается от алгоритма любого иного текстового поиска. Обсуждение задач такого рода мы не выделяем в отдельный учебный модуль, лишь акцентируем внимание студентов на уже известных им положениях:

- названия веществ многовариантны (например, *дигидрофосфат натрия*, *натрий дигидроортофосфат*, *однозамещенный фосфат натрия* и т. п.);
- полнота извлечения информации обеспечивается учетом всех синонимов и морфологических форм терминов в запросе.

Для того чтобы устранить проблемы, вызванные многовариантностью, веществам в больших базах данных присваивают регистрационные номера. Ссылки на такие номера встречаются и за пределами исходного массива, поэтому пользователю целесообразно иметь некоторое представление о них хотя бы на уровне распознавания. Формат отображения регистрационного номера и местонахождение соответствующей базы данных полезно знать, например, для *UN ID* (международный код вещества, представляющего

опасность при транспортировке), *RTECS* # (код в базе данных токсичности веществ), *EC* # (код химического товара в странах ЕС) – велика вероятность, что выпускник химфака может столкнуться с ними в своей профессиональной деятельности.

Более пристальному рассмотрению подлежат *CASRN* – регистрационные номера *Chemical Abstracts Service,* – поскольку они широко распространены в литературе и во многих информационных массивах де-факто выполняют функции стандартных идентификаторов химических веществ [2].

Практически все справочные базы данных имеют инструменты для целенаправленного и эффективного обнаружения сведений о веществе по его известному коду *CASRN*. Этот же код с успехом можно использовать в составе текстового запроса *Google* для поиска в открытом Интернете.

В соответствующем разделе учебного курса мы уделяем большее внимание не самому поиску по *CASRN* – он вполне стандартен и интуитивно понятен, – а тем особенностям регистрационной системы *CAS*, которые рядовой пользователь нередко не замечает, но которые существенны для планирования и адекватной оценки результатов работы.

Правильное соотнесение *CASRN* и вещества – задача не такая простая, какой она кажется при беглом взгляде на проблему. *CAS* регистрирует химические объекты, то есть не только химические вещества в строгом понимании этого термина (например, три разных регистрационных номера имеют цис-1,2-дихлорэтен, транс-1,2-дихлорэтен, а также 1,2-дихлорэтен без указания симметрии).

Упомянутую выше особенность студенты исследуют в ходе практикума. Они должны самостоятельно обнаружить, что в каталоге *Sigma-Aldrich* химическое вещество *этан* представлено почти десятком реактивов (это, кроме $C_2H_6$, этаны дейтерированные или содержащие $^{13}C$) и каждый из них характеризуется своим *CASRN*. Последующий анализ в иных базах данных показывает, что *CAS* присвоил уникальные регистрационные номера и некоторым этансодержащим смесям (углеводороды $C_1$-$C_2$, $C_2$-$C_3$ и др.).

В дискуссии по результатам проведенного эксперимента студенты формулируют предложения по рациональному использованию *CASRN* в информационном поиске.

Второй проблемой является степень достоверности тех кодов *CASRN*, с которыми сталкивается пользователь.

Исчерпывающий список всех *CASRN* содержится в *CAS REGISTRY* – платной базе данных. *Chemical Abstracts Service* разрешает некоммерческое использование *CASRN* – но при условии, что в информационный источник будет включено менее 10 тыс. регистрационных номеров. В результате такой политики в печатной литературе (особенно в каталогах реактивов) и в онлайновых справочных базах данных накопилось значительное количество кодов *CASRN*, однако это лишь мизерная доля той информации, которой владеет *CAS*.

Для значительной части научной аудитории доступ к платному первоисточнику (*CAS REGISTRY*) затруднен, что стимулирует обращение ко вторичным документам. В результате в онлайновой литературе из одного документа в другой копируются коды с опечатками и отмененные коды, и нередко встречается некорректное соотнесение *CASRN* и вещества.

Трудности возникают у авторов документов и у составителей справочных баз данных, решивших дополнить свои материалы кодами *CASRN*. Для правильного выбора кода необходимо профессионально ориентироваться в регистрационной системе и иметь доступ к первоисточнику *CAS REGISTRY*. Как показывает анализ качества бесплатных веб-ресурсов, эти условия выполняются далеко не всегда.

Трудности возникают у пользователей, намеревающихся найти в Интернете информацию о веществе по коду *CASRN*. Пользователь должен знать «правильный» код, да к тому же еще и уметь предвидеть, какие «частично правильные» коды могут появиться в онлайновой литературе (присвоение кристаллогидрату регистрационного номера безводной соли и т. п.).

Результат спонтанного накапливания ошибок мы демонстрируем на примере перспективного и во многих отношениях весьма полезного ресурса *ChemSpider* (*www.chemspider.com*). Этот прежде частный сайт перешел в собственность *Royal Society of Chemistry*, и теперь он превращается в крупнейший онлайновый центр справочной, структурной и спектральной информации.

По своему строению *ChemSpider* – это база данных, наполняемая из многочисленных коммерческих и некоммерческих источников, плюс поисковая система, способная обрабатывать не только текстовые, но и структурные запросы. В функционировании *ChemSpider* присутствуют элементы *Web 2.0*, а именно самоархивирование и открытое рецензирование опубликованного материала.

Выполняя упражнение, студенты извлекают блок идентификаторов этана; блок содержит около пяти *CASRN*. Как и полагается в ресурсах категории *Web 2.0*, в списке отражены результаты работы онлайновых рецензентов (*Validated by Experts*, *Validated by Users*, *Non-Validated* и др.).

Тщательный анализ с использованием авторитетных справочников подводит учащихся к парадоксальному выводу: *CASRN* индивидуального вещества этана здесь фигурирует как «не подтвержденный», а коды смесей, всего лишь содержащих этан в своем составе, – как «утвержденные экспертами».

Данное исследование позволяет еще раз привлечь внимание пользователей к принципам онлайновой работы:

- источником справочной информации о веществе должна служить только редактируемая литература (в том числе, редактируемые научные базы данных);
- репозитории категории *Web 2.0* следует рассматривать как полезные, но вспомогательные ресурсы, содержащие сведения неопределенной точности.

Отметим, что *Chemical Abstracts Service* разместил официальную выборку из *CAS REGISTRY* на сайте *Common Chemistry* (*www.commonchemistry.org*), а онлайновое сообщество совместно с *CAS* провело сверку *CASRN* для 9 тыс. веществ, присутствующих на страницах *Wikipedia*. Теперь эти регистрационные номера – чуть ли не единственный элемент Википедии, имеющий официальное подтверждение своей достоверности.

Наиболее важным идентификатором вещества является химическая формула.

Современные текстовые поисковые программы способны обнаруживать документы, содержащие заданные линейные формулы (брутто-, рациональные). Эти программы распознают, в том числе, записи с нижними и верхними числовыми индексами – но не всегда корректно.

В отдельных базах данных допускается та или иная степень неопределенности в конструировании запроса вплоть до простого перечисления символов химических элементов без указания количества атомов в молекуле или формульной единице.

При рассмотрении алгоритма поиска по заданной одномерной химической формуле дополнительному обсуждению на занятиях подлежат:

- варианты отображения числовых индексов при использовании линейной формулы в запросе;
- правила записи брутто-формулы по системе Хилла и порядок расположения веществ в формульном указателе.

В Интернете постоянно растет объем материала, характеризующего топологию органических и неорганических веществ, причем возможность использования *2D*-структур в качестве запроса перестала быть привилегией коммерческих баз данных.

Приемы и инструменты обнаружения информации по двумерной формуле химического вещества подлежат детальному изучению в практикуме.

Как правило, в структурных базах данных реализуется три алгоритма работы поисковых программ:

- Поиск по структуре – *Exact (Structure) Search*. (Извлекаются вещества, полностью соответствующие структурной формуле, указанной в запросе).
- Поиск по подструктуре (субструктуре) – *Substructure Search*. (Извлекаются вещества, в структуре которых есть участок, целиком совпадающий с остовом запроса).
- Поиск по подобию – *Similarity Search*. (Извлекаются вещества, имеющие такие же структурные фрагменты, какие есть в запросе. На поисковом бланке обычно можно задать минимально допустимую степень соответствия, в %. Для количественной характеристики степени соответствия часто используется коэффициент Танимото).

Поиск по структуре и субструктуре – это материал, обязательный для изучения. К поиску по подобию химики прибегают главным образом при решении специальных задач, поэтому мы только знакомим студентов с этим алгоритмом.

В структурных базах данных наиболее распространенным способом формулирования запроса является конструирование двумерной формулы в поле апплета [3].

*Апплет* – это маленькая исполняемая программа, которая загружается вместе с веб-страницей, содержащей поисковый бланк. Приемы работы с апплетом весьма схожи с приемами работы с обычным молекулярным редактором, таким как *ISIS/Draw*; при минимуме навыков они оказываются интуитивно понятными. Основная задача апплета – правильно отобразить молекулярный граф, на основании которого должен проводиться поиск. Красота формируемого изображения не имеет значения в поисковом процессе, поэтому у таких апплетов нередко отсутствуют дизайнерские функции (выравнивание длин связей, валентных углов, поворот в плоскости листа и т. п.). Инструментарий апплета может ограничиваться

управляющими командами (стереть, выделить) и кнопками химических символов, химических связей, простейших структурных фрагментов.

Из имеющихся в *WWW* инструментов подобного назначения апплет информационного центра *NIST Chemistry WebBook*[1] характеризуется, по-видимому, самым простым строением; именно здесь студенты осваивают основные приемы редактирования двумерных формул и ведения структурного поиска.

Процесс «рисования» не должен восприниматься пользователем как самоцель и не должен отвлекать от главного – анализа извлекаемой информации. Такое условие выполнимо, если пользователь научится распознавать однотипные элементы у разных апплетов, а навыки выполнения вспомогательных операций будут доведены у него до автоматизма.

Выполняя тренировочные упражнения, на первом этапе студенты осваивают стандартные функции апплета как молекулярного редактора: режимы работы (выделение, рисование, удаление, обращение к шаблонам); отображение атома и химической связи; замена свойства атома, изменение порядка связи; ввод данных с клавиатуры.

На втором этапе рассматриваются специфические детали, такие как отображение атомов водорода в явной и неявной формах и влияние вида молекулярного объекта на ход субструктурного поиска.

Проверочные задания включают в себя поиск по структуре (например, «Определить стандартную энтальпию образования вещества *X*») и субструктурный поиск (например, «Определить температуры плавления бромопроизводных вещества *Y*»); здесь *X* и *Y* – двумерные формулы.

Работа завершается кратким исследованием: студенты, варьируя исходную структуру, определяют, какие именно атомы и связи данная программа считает остовом в ходе субструктурного поиска.

На большой группе сайтов формулирование структурных запросов осуществляется с помощью апплета *JME*. В нашем практикуме приемы

[1] *webbook.nist.gov/chemistry*

работы с *JME* рассматриваются на примере каталога реактивов компании *Sigma-Aldrich*[2]. Интерфейс апплета прост настолько, что мы предлагаем студентам самостоятельно исследовать его функционирование методом проб и ошибок при минимальной помощи преподавателя.

В контрольное задание по разделу включен субструктурный поиск, анализ списка обнаруженных результатов, извлечение карточки заданного вещества, а также изучение схемы предоставления информации о ценах на реактив в типичном онлайновом каталоге.

Еще один апплет – *MarvinSketc*h – в последнее время получает все большее распространение в Интернете.

В отличие от апплетов, рассмотренных выше, в интерфейсе *MarvinSketch* присутствуют пункты меню, отдельные команды которого дублируются кнопками, а также пункты меню и инструменты, которые не используются в конкретной базе данных.

Студентам предлагается экспериментально выяснить приемы создания и редактирования двумерных формул в соответствии с планом, изложенным в методическом руководстве. По окончании подготовки они переходят к информационному поиску, который проводится на сайте *chemicalize.org* (*www.chemicalize.org*).

*Chemicalize.org* – поисковая система нового типа: индексируя научный текст, она распознает химически значимые фрагменты; более того, система семантически обогащает документ – дополняет его структурными формулами и тематическими ссылками.

В этом учебном модуле на стадии контроля студент должен с помощью *chemicalize.org* найти и извлечь текст патента, содержащего сведения о заданном веществе, проанализировать гипертекстовую структуру *HTML*-документа, обнаружить и исследовать элементы его семантического обогащения (и таким образом ознакомиться с идеями, которые, как полагают, будут реализованы в вебе нового поколения – *Web 3.0*).

[2] *www.sigmaaldrich.com/catalog/AdvancedSearchPage.do*

Отдельную категорию идентификаторов химических веществ составляют линейные нотации *SMILES* и *InChI*; в традиционной печатной литературе они встречаются крайне редко, но являются вполне рядовыми элементами электронных баз данных. Линейная нотация отображает молекулярный граф химического объекта в виде строки буквенно-цифровых символов и может использоваться как основной компонент запроса в информационном поиске.

Система машинной обработки химических структур *SMILES* применяется более двадцати лет, в том числе, и для обмена данными между разнотипными информационными массивами. Коды *SMILES*, хотя и ориентированы в основном на компьютер, легко прочитываются пользователем, имеющим некоторое представление об их синтаксисе (например, CC(=O)O — уксусная кислота) [4]. Для успешной работы химику полезно знать принятые здесь основные правила отображения атомов, ковалентных химических связей, ионов и ионных соединений, алифатических и ароматических циклов, пространственных изомеров, автономных частиц, схем химических реакций.

Выполняя тренировочные упражнения, студенты составляют коды *SMILES* для серии структур и затем самостоятельно проверяют их правильность с помощью удобного и быстрого конвертора, имеющегося в каталоге *Sigma-Aldrich*.

Контрольное задание содержит требование обнаружить, используя *SMILES* в качестве запроса, указанное вещество и его заданные свойства в базе данных *ChemSpider*.

В ходе обсуждения результатов работы студенты должны сформулировать, что использование *SMILES* в запросе целесообразно для:

- поиска в текстовых базах данных (где структурную формулу невозможно применить в принципе);

- поиска в структурных базах данных, если код прост в написании (например, легче набрать на клавиатуре строку CC(N)C(O)=O, чем рисовать двумерную структуру аланина);
- поиска в структурных базах данных, если готовый код можно скопировать из другого источника, например, из молекулярного редактора (из-за несовместимости форматов молекулярных редакторов далеко не всегда возможен перенос двумерной формулы; для буквенно-цифровой строки *SMILES* проблема несовместимости отсутствует).

Международный текстовый идентификатор *InChI*, как предполагается, должен стать стандартной формой представления информации о структуре вещества [5]. *InChI* (а также *InChIKey* как аналог *DOI* для молекулярных структур) генерируется и интерпретируется компьютером; пользователю для осмысленной работы достаточно иметь лишь начальные сведения о формате и структуре кода.

Информационный поиск по запросу, состоящему из *InChI* или *InChIKey*, пока что не обладает какими-то особыми преимуществами по сравнению с другими методами. Причина кроется в том, что эти коды – явление достаточно новое и, хотя горячо рекламируемое своими сторонниками, не получившее повсеместного распространения.

В практикуме мы знакомим студентов с тремя рекомендованными ИЮПАК генераторами кодов *InChI* (*ChemSketch*, *ChemSpider* и *PubChem Sketcher*) и проводим краткое исследование возможности использования *InChI* в запросе поисковой системы *Google*.

Навыки работы с различными типами идентификаторов химических веществ совершенствуются на последующих занятиях учебного курса.

## ЛИТЕРАТУРА

**1.** Рагойша А.А. Информационные технологии в химии: учебные материалы практикума. Ч. 2. Химическая структура. – Минск: БГУ, 2010. – URL: http://www.abc.chemistry.bsu.by/2/.

**2.** Рагойша А.А. CAS Registry Number и справочник Common Chemistry. – Минск: БГУ, 2009. – URL: http://www.abc.chemistry.bsu.by/bulchinf/2009_1_6-8.pdf.

**3.** Ertl P. Molecular structure input on the web. Journal of Cheminformatics, 2010, 2:1. – URL: http://www.jcheminf.com/content/2/1/1.

**4.** Daylight Theory Manual. Daylight Chemical Information Systems, Inc., 2008. – URL: http://www.daylight.com/dayhtml/doc/theory/index.html.

**5.** Heller S.R., McNaught A.D. The IUPAC International Chemical Identifier (InChI). Chemistry International, 2009, 31, 1, p. 7 – 9.