ПРОБЛЕМА КАЖУЩЕЙСЯ ОЧЕВИДНОСТИ В ОНЛАЙНОВОМ ИНФОРМАЦИОННОМ ПОИСКЕ

А. А. Рагойша
Белорусский государственный университет
ragoisha@bsu.by

A PROBLEM OF ILLUSIVE EVIDENCE IN ONLINE INFORMATION RETRIEVAL

A. Rahoisha

Previous experience may encourage students to follow ineffective strategy when they proceed from general to specialized information retrieval; a common problem is their belief that Google is the only suitable tool to deal with any kind of search tasks. A special training should be undertaken to persuade the students to perform actions that seem obvious to the scientist. The issues to be discussed in class include uncertain quality of retrieved documents, simplifying procedures of evaluation of information, shortcomings of ranking based on the popularity of the resource, the need for critical analysis of data that is published in edited sources.

ЗНАНИЕ: Google — это:

- огромная база данных,
- эффективный алгоритм ранжирования,
- безупречная техническая поддержка

(высокая скорость работы, бесперебойность).

ЗНАНИЕ: Интернет анархичен.

Google — зеркало Интернета (не кривое!).

Google легко находит.

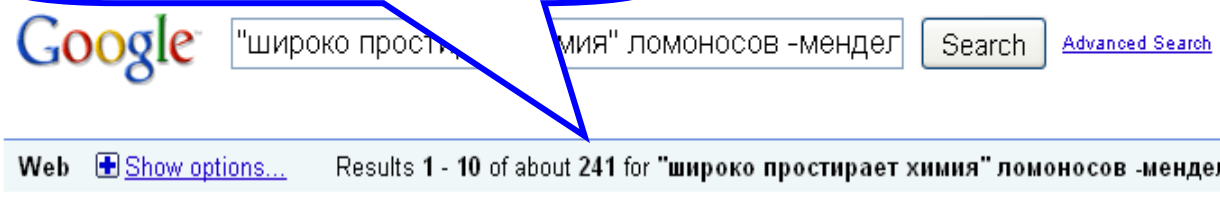
Человек тщательно перепроверяет найденное.

ВЕРА:
Google ЗНАЕТ ВСЕ

Пример. *Кто автор фразы:*

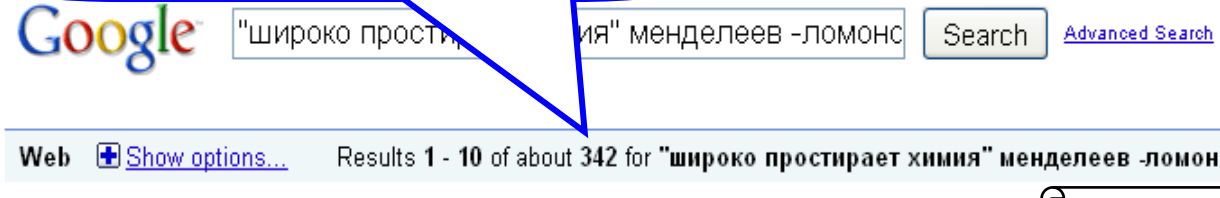
"Широко простирает химия руки свои в дела человеческие"

241 документ — за Ломоносова.



manaev: "**Широко простирает химия** руки свои в дел - [Translate this page]
"**Широко простирает химия** руки свои в дела человеческие". - **Ломоносов**. (Post a new comment). [info] lee_van_cleef 2006-07-11 07:08 pm UTC (link) **...**
manaev.livejournal.com/634362.html

342 документа — за Менделеева?!.

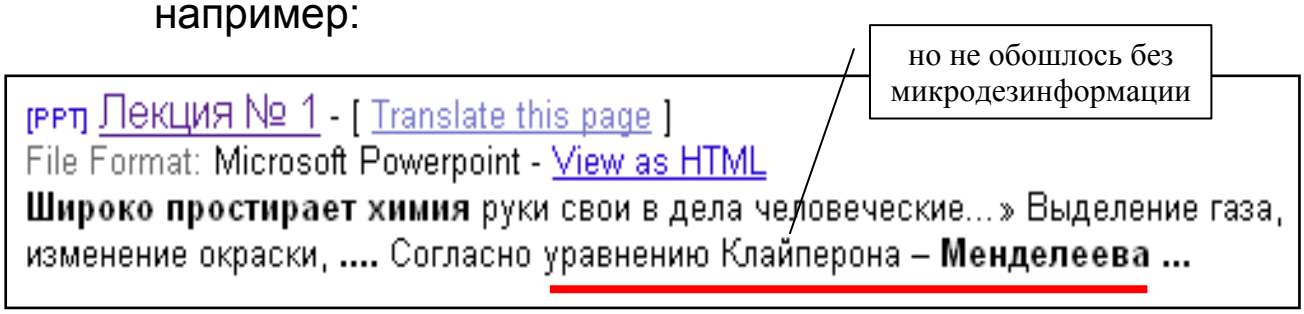


Химия и жизнь (Шимшон Шиманский) / философия / Проза.ру ... - [Translate this pa
6 фев 2010 **...** "**Широко простирает химия** руки свои в дела человеческие". (Д.И.**Менделеев**) Есть большая группа элементов в Периодической таблице **Менделеева**, **...**
www.proza.ru/2010/02/06/1145 - Cached

Правда крупно проиграла

Результаты поиска улучшаются при тщательном рассмотрении:

(1) *n*-ная доля из обнаруженных выше 342 документов все же *не* приписывает Д.И.Менделееву авторства, например:

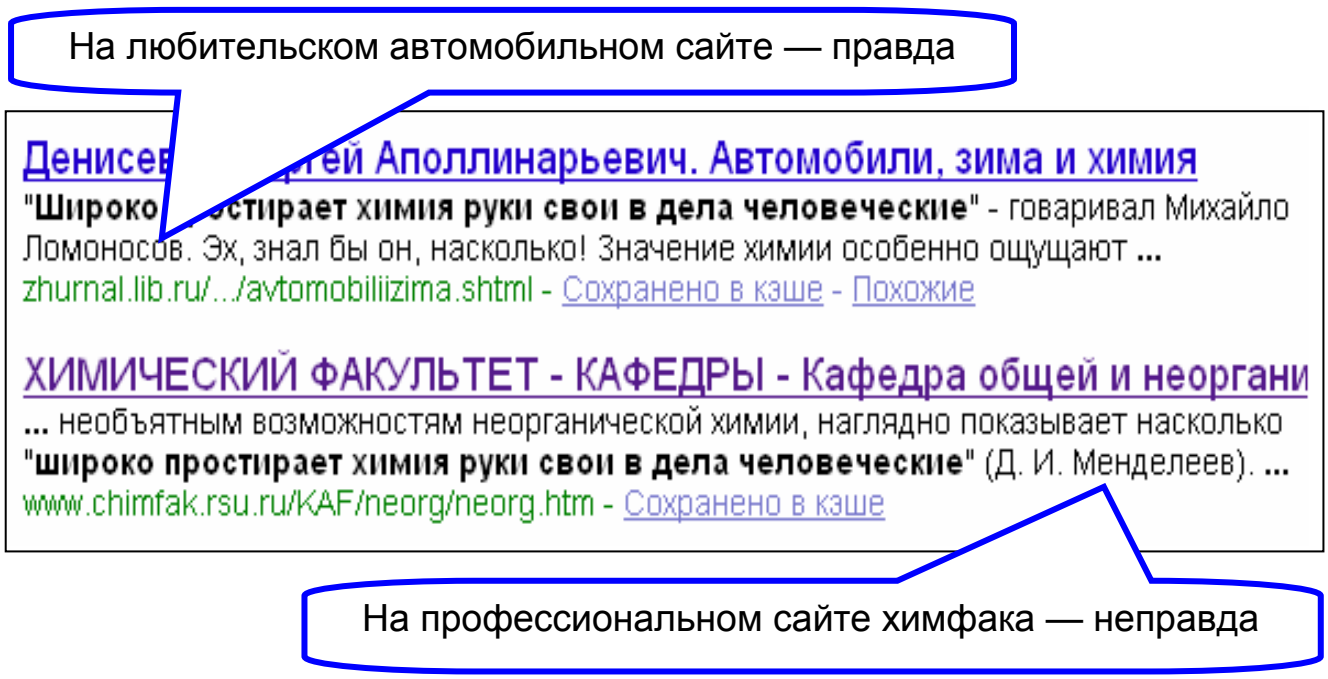


(2) Если запрос состоит только из фразы "широко простирает химия...", в начале извлеченного списка свидетельств в пользу Ломоносова больше, чем в пользу Менделеева.
Здесь заслуга Google очевидна.

Критерий оценки достоверности документа: **КТО АВТОР**?

Учитывать — ДА, бездумно верить — НЕТ.

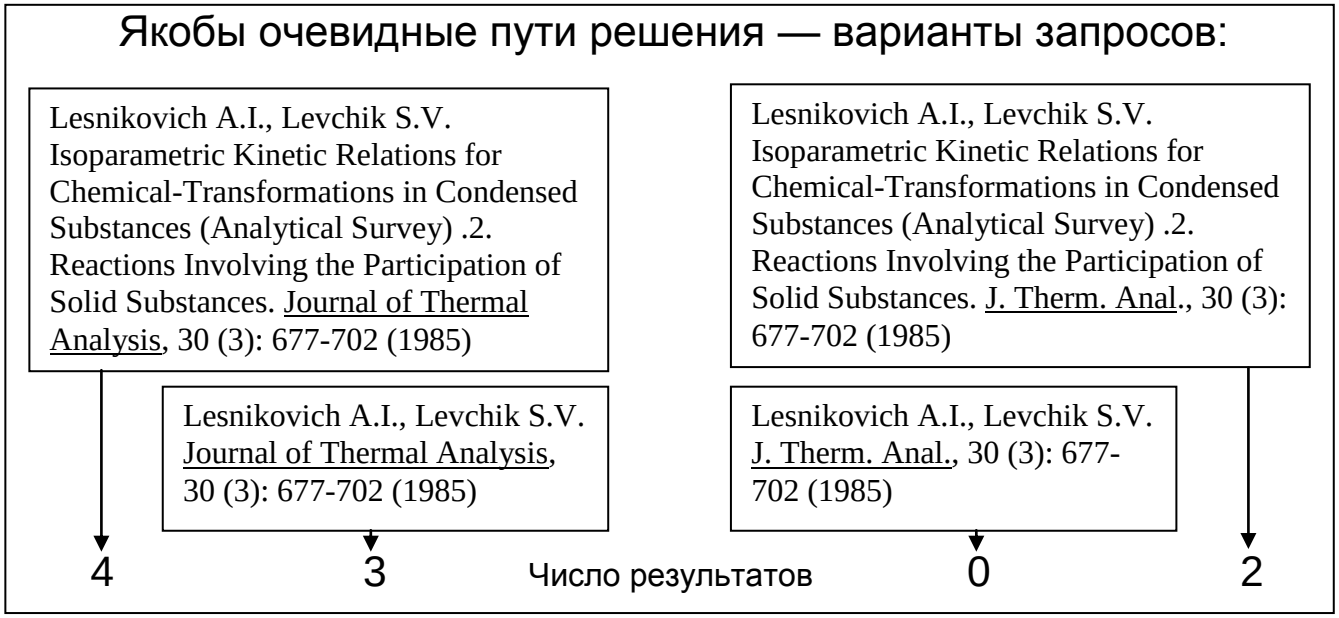
Пример — две соседние записи в списке результатов:



Google в библиографическом поиске

Задача. Известно полное библиографическое описание статьи.

Найти первоисточник.



Что же в списке результатов поиска ?

Первоисточник отсутствует.

Обнаруженные документы всего лишь цитируют эту статью.

Переформулировать запрос?

Перебор вариантов → Затраты усилий с неясной перспективой.

Как действовать?

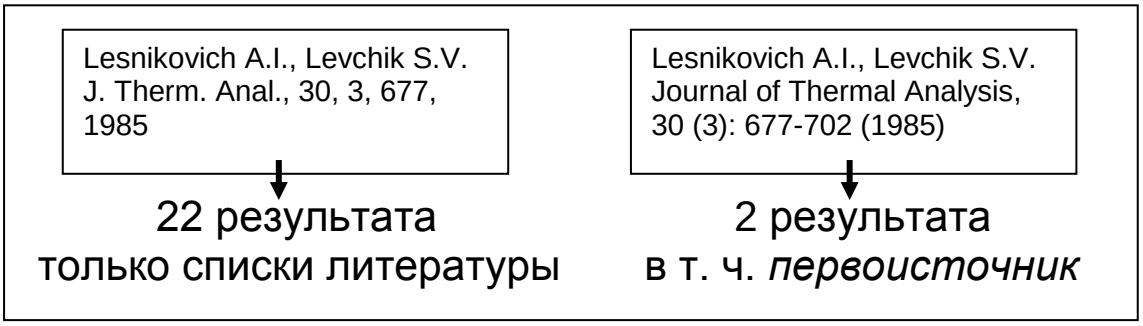
1. *Осознать*: не мы первые попали в Интернет.
   Задолго до нас профессионалы разрабатывали вспомогательные поисковые инструменты и алгоритмы поиска, отбирали материал для специализированных баз данных, формировали тематические метасайты, рецензировали научные работы.
2. *Не изобретать велосипед*. Не пытаться начинать с нуля решение любой поисковой задачи. Максимально использовать наработки профессионалов, по крайней мере, на стадии первичного отбора материала.
   Одно из следствий: стремиться использовать (только) рецензируемые информационные источники.
3. Сэкономленные умственные усилия направить на **анализ качества и содержания** извлеченных документов.

Пример: библиографический поиск.

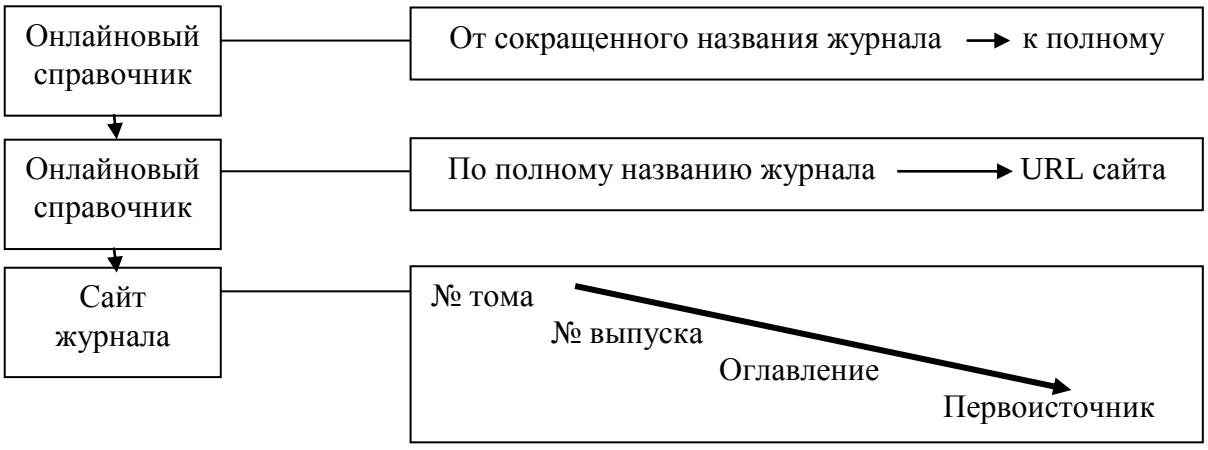
Поиск первоисточника по известному библиографическому описанию следовало бы проводить в библиографической базе данных — полнотекстовой либо содержащей URL (DOI) статей.

Эффективные инструменты (доступность ограниченна):
Scopus, ISI Web of Knowledge.

Удовлетворительный доступный инструмент: Google Scholar.



Универсальный альтернативный алгоритм:
в центре внимания — сайт журнала.



В рутинной работе алгоритм практически всегда приводит к успеху, но **нелюбим студентами** за его многостадийность.

Достоверность и авторитетность автора (издателя)

Бесплатных хороших ресурсов — все больше.

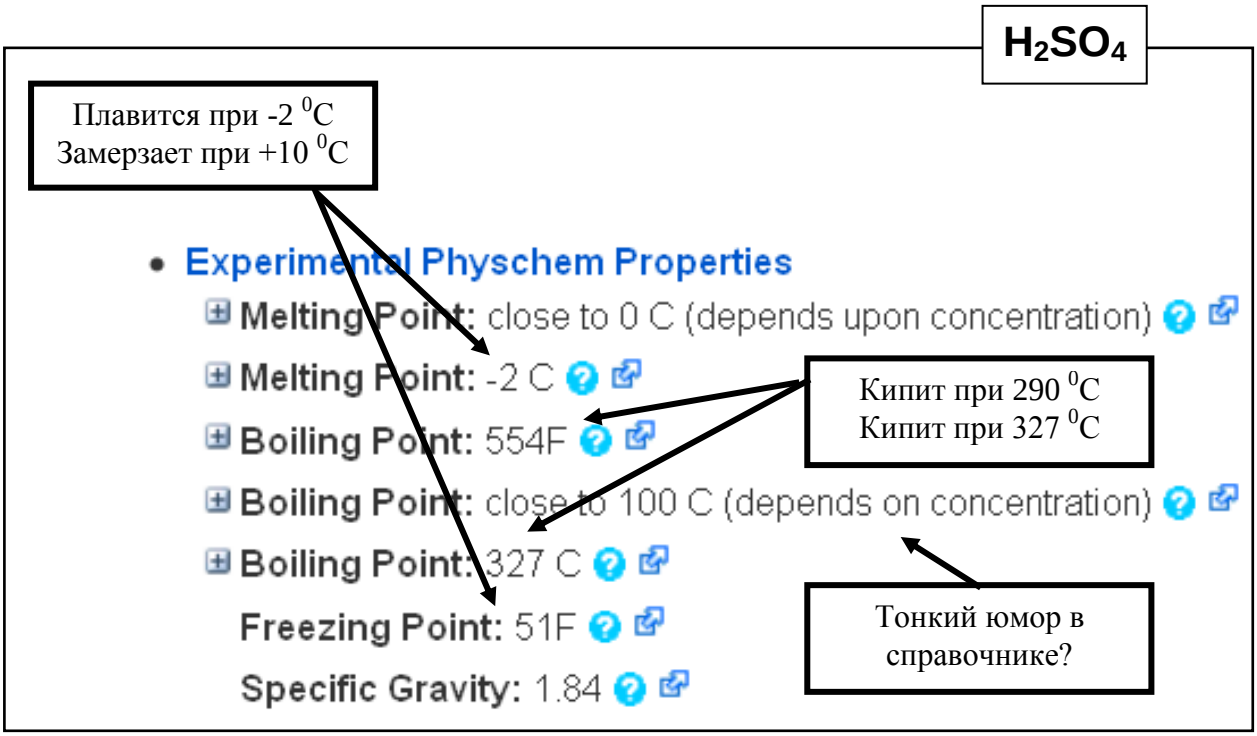
Но ... беды роста.

- повальный переход от осмысленного ручного труда к автоматизации;
- гонка за количеством методом *Copy&Paste*;
- пренебрежительное отношение к редакторской и корректорской работе;
- безответственность:
  принцип "*as is*" — общепринятая норма.

Грань между категориями "редактируемая база данных" и "репозиторий" размывается. Размываются границы между достоверным, сомнительным и недостоверным.

Иллюстрации из очень хороших баз данных.

Пример 1. **Физические свойства H_2SO_4 в *ChemSpider* (экспериментальные данные):**

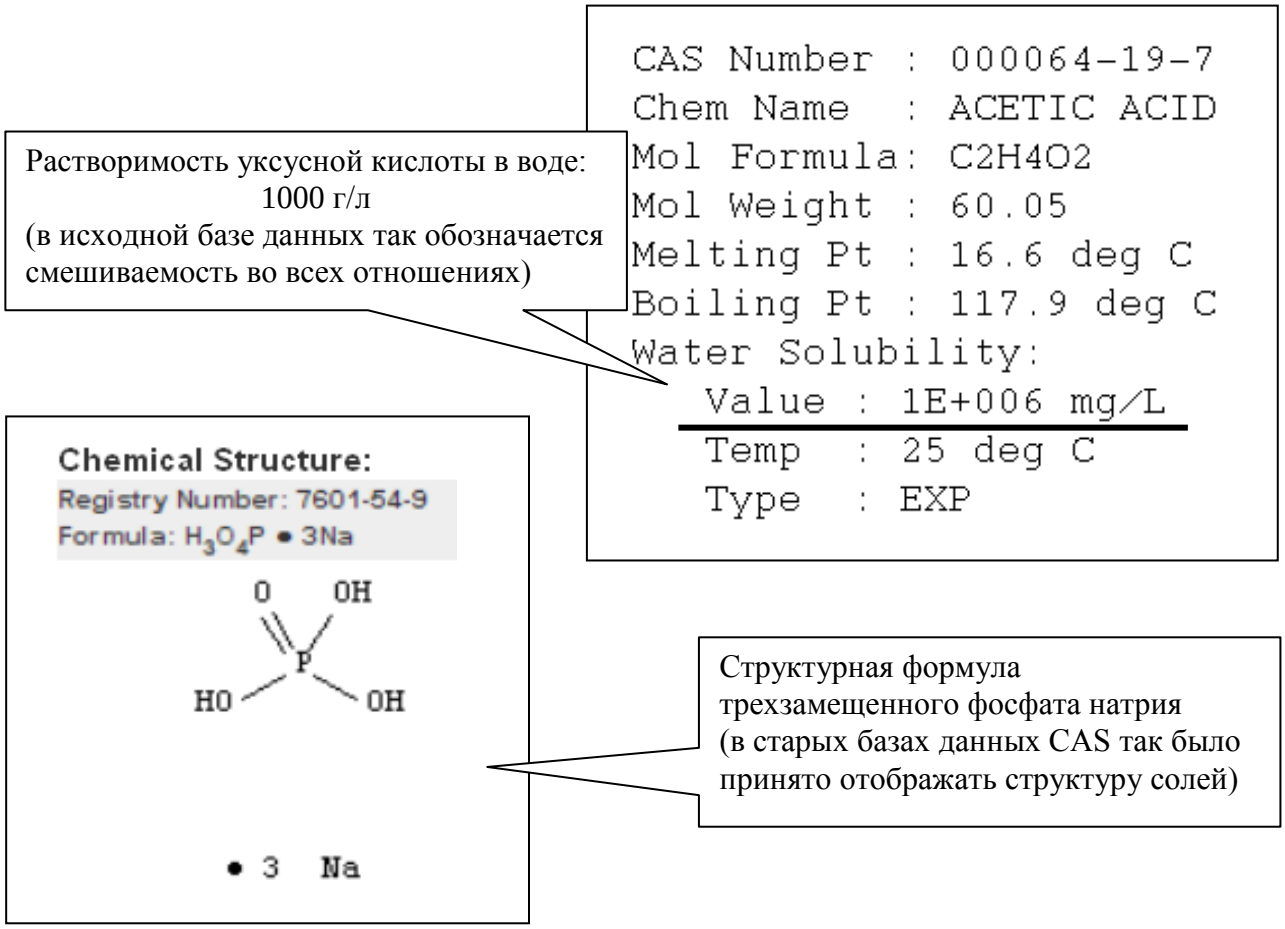


ChemSpider позиционирует себя как ведущий общемировой центр в области структурной органической химии.
С 2009 г. — это собственность *Royal Society of Chemistry*. Авторитетность *RSC* не вызывает сомнений, и вполне закономерно, что доверие пользователя автоматически распространяется на весь *ChemSpider*. Как показывает практика, такой подход не всегда оправдан.

В приведенном выше примере пользователь должен *догадаться*, что перед ним не рекомендованные составителем, а механически собранные из Интернета числа. Проблема оценки их достоверности *перекладывается* на самого пользователя.

Пример 2. **Артефакты, в WWW превращающиеся в факты**

В первоисточнике могут встречаться обозначения, условность которых оговаривается в поясняющих документах.
После копирования в иные базы данных эти элементы приобретают для непосвященного пользователя черты реальных параметров.



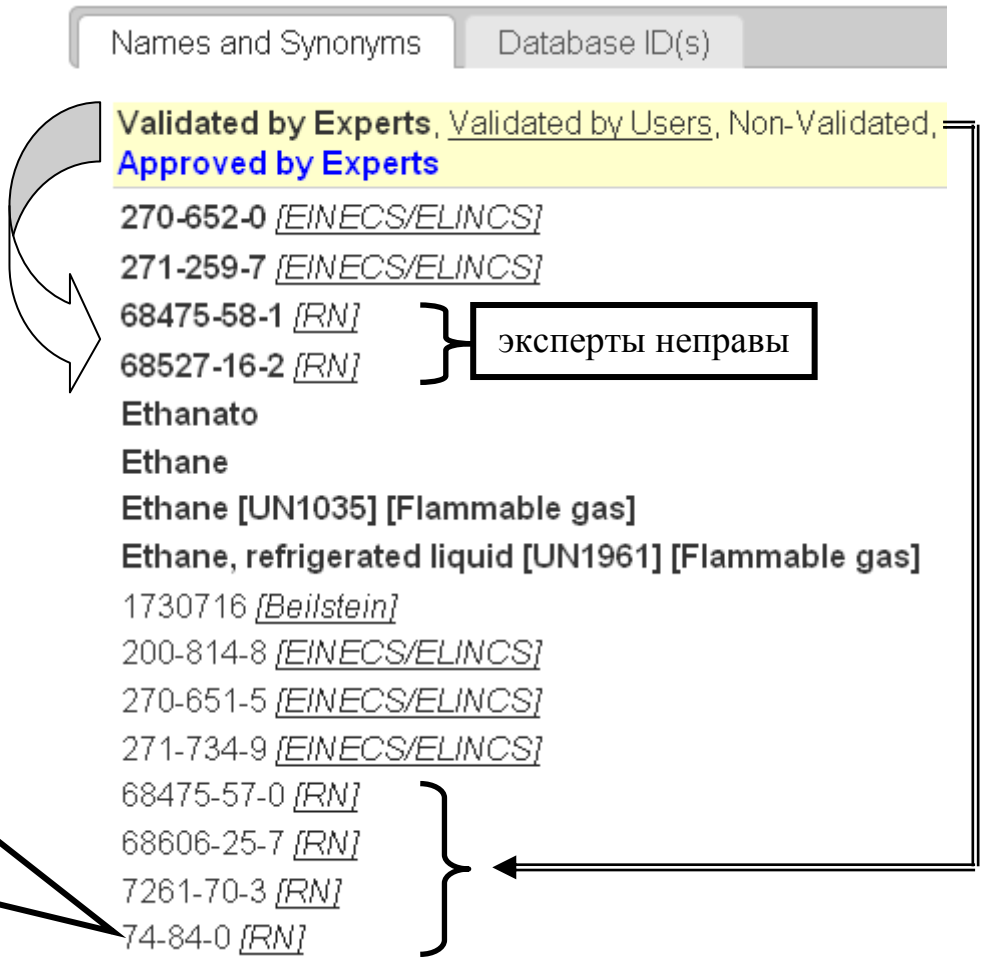
Достоверность и онлайновое рецензирование

Надежды, что коллективный контроль качества онлайновой информации окажется эффективным, за редким исключением, не оправдались.

При оценке достоверности документа результат коллективной экспертизы следует считать не решающим, а совещательным.

Пример (ChemSpider).

Каков CAS Registry Number этана?



Правильный ответ находится в конце списка

Журнал, практикующий онлайновое рецензирование статей и *считающийся* рецензируемым, на практике может содержать *нерецензированные* публикации.