

Информационные технологии в химии

**Александр Антонович
Рагойша**

**Кафедра общей химии и методики
преподавания химии
к. 501-а**

1-й семестр

**Поиск химической информации
в онлайн-овых текстовых базах данных**

58 часов, из них аудиторных – 38:

8 – лекции

24 – практикум

6 – КСР

Зачет (1,5 зач. ед.)

ЛИТЕРАТУРА

- А. А. Рагойша. Поиск химической информации в Интернете. Поисковые системы и тематические каталоги: Учеб. пособие для студентов хим. фак. – Мн.: БГУ, 2003.
- А. А. Рагойша. Поиск химической информации в Интернете: научные публикации : учеб. пособие для студентов хим. фак. спец. 1-31 05 01. – Мн.: БГУ, 2007.
- В. М. Потапов, Э. К. Кочетова. Химическая информация. Где и как искать химику нужные сведения. – М.: Химия, 1988.
- Рагойша, А. А. [Текстовый поиск научной химической информации в Интернете] : практикум по курсу "Информационные технологии в химии" для студентов спец. 1-31 05 01 Химия (по направлениям) — Мн.: БГУ, 2012.
<http://elib.bsu.by/handle/123456789/14599>
- А. А. Рагойша. Азбука веб-поиска для химиков. – Минск, БГУ, 1999-2014. <http://www.abc.chemistry.bsu.by>.

Курс "Информационные технологии в химии"

1 семестр (н.-пр., пед.)

Лекции: 1-2 3-4 5

1 семестр (фарм.)

Лекции: 1 2 3

Архив лекций

Все учебные программы

Практикум

1. Поиск химической информации

2. Патентные базы данных

3. Практикум. Часть 2.

Бюллетень химической информации

CrossMark — индикатор степени достоверности научной статьи (PDF 134 KB)

Бюллетени 2009-10

Архив



ABC Chemistry: Бесплатные полнотекстовые научные журналы по химии

1. Каталог постоянно доступных химических журналов
2. Информация о временно доступных химических журналах

Навуковыя хімічныя часопісы - праз сетку БДУ

Навуковыя часопісы - праз сеткі бібліятэк Беларусі

ABC Chemistry: Free Full-Text Journals in Chemistry

- A. Directory of permanently available chemical journals
- B. Trials and temporarily available chemical journals



А. А. Рагойша.
Поиск химической информации в Интернете. Поисковые системы и тематические каталоги. Минск, 2003
PDF, 647 KB



А. А. Рагойша.
Интернет для начинающих и не только... Минск, 2004.
PDF, 873 KB



А. А. Рагойша.
Поиск химической информации в Интернете: Научные публикации. Минск, 2007
PDF, 770 KB



А. А. Рагойша.
Текстовый поиск научной химической информации в Интернете : руководство к практикуму. Минск, 2011
PDF, 669 KB

Терминология

WWW

- Интернет

— (*inter* — меж- + *net* — сеть) —
сеть, объединяющая много компьютерных сетей.

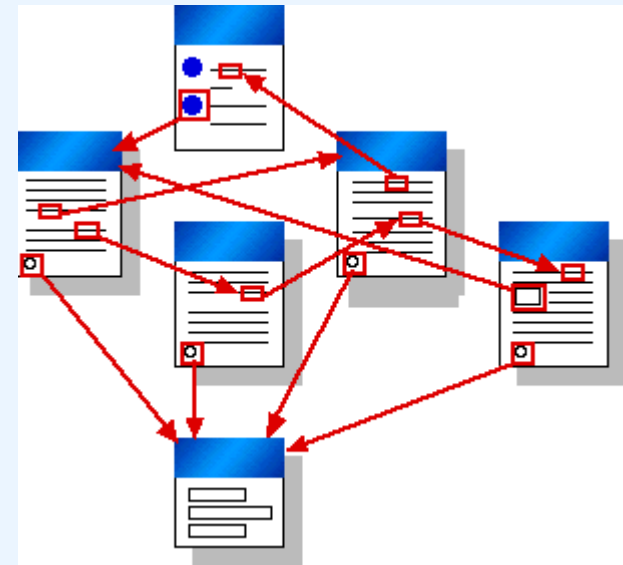
- World Wide Web

(WWW, Web, W3, Всемирная паутина, веб) —
система взаимосвязанных между собой документов,
доступных через Интернет.

Документ — любой целостный автономный
информационный массив, не только текстовый, но и,
например, видео-, аудио- и т. д.

Гипертекст

- *Протокол* — набор правил.
- **HTTP (Hypertext Transfer Protocol)** — протокол передачи гипертекста.
- **Гипертекст** — «текст ветвящийся или выполняющий действия по запросу» (Тед Нельсон, 1965).
- **Гиперссылка (ссылка, link)** – часть гипертекстового документа, указывающая на другую часть этого документа или на другой документ.



Домен

- **IP-адрес** —
числовой идентификатор компьютера(ов) в сети.

Пример: **217.21.43.222**

- **Доменное имя** —
буквенно-числовой идентификатор узлов сети и ресурсов, расположенных на узлах.

Иерархическая структура

Примеры: **www.abc.chemistry.bsu.by**
www.cam.ac.uk
www.google.com

Домен верхнего уровня

- **Общий домен верхнего уровня**
без регистрационных ограничений
com, net, org, info
с ограничениями («спонсируемые»)
gov, int, mil, edu, museum, biz, ...
- **Национальный домен верхнего уровня**
by, uk, ru, de, ..., eu
tv, la (... и за пределами страны)
рф

Структура

- **Сайт** (веб-сайт, **website**, ...) — информационный массив, находящийся на сервере и доступный внешним пользователям.

Единый стиль

Структура может быть иерархичной

- **Веб-страница** (страница, **webpage**, **page**) — документ, который можно получить в ходе одного обращения к серверу.

Веб-страницы: статические, динамические

Адрес

- Адрес (URL, Uniform Resource Locator) - стандартизированный указатель местонахождения информации и способа ее получения.

<http://www.abc.chemistry.bsu.by/current/bdu.htm>

<http://www.bl.uk/eresources/jnls/ejournals.html#free>

<http://www.bsu.by/ru/main.aspx?guid=4681>

<http://scout-unimib.cilea.it/links/SPT-->

[FullRecord.php?ResourceId=491&PHPSESSID=d666f9f88fe19ef1](http://scout-unimib.cilea.it/links/SPT--FullRecord.php?ResourceId=491&PHPSESSID=d666f9f88fe19ef1)

<http://ru.wikipedia.org/wiki/%D0%91%D0%93%D0%A3>

(<http://ru.wikipedia.org/wiki/БГУ>)

<ftp://ftp.netscape.com/robots.txt>

Сайт

- Главная страница (Первая, Home Page, Main Page, ...) —
титульная веб-страница информационного массива.

страница по умолчанию (default page)

www.abc.chemistry.bsu.by

<http://www.abc.chemistry.bsu.by/>

<http://www.abc.chemistry.bsu.by/default.htm>

<http://www.12345.org/>

[default.htm](#) [index.htm](#)

[default.html](#) [index.html](#) [index.php](#)

Исчезла страница?

www.1abc.2def.org/mmm/nnn/ppp.htm?id=222

www.1abc.2def.org/mmm/nnn/ppp.htm

www.1abc.2def.org/mmm/nnn/

www.1abc.2def.org/mmm/

www.1abc.2def.org/

1abc.2def.org/

www.2def.org/

Поисковая система

(Search engine)

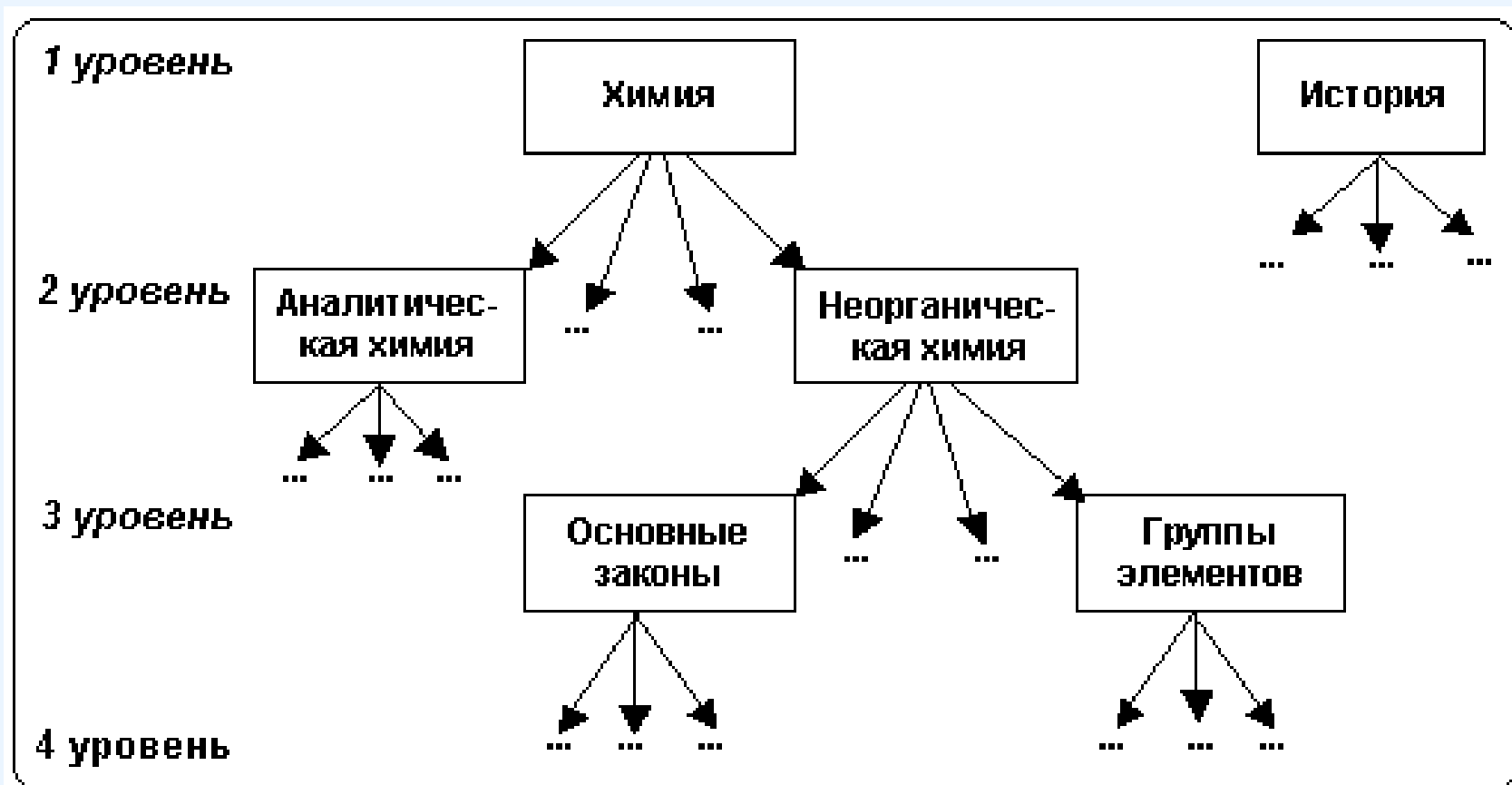
- робот (паук)
- индекс (база данных)
- поисковая программа, веб-интерфейс

Универсальные поисковые системы:

Google, Yahoo!, Bing, Яндекс, ...

Специализированные (вертикальный поиск)

Тематический каталог



Каталог (Directory)

Раздел (Category)

Еще указатели веб-ресурсов:

- **Метапоисковая система**
использует индексы нескольких иных поисковых систем
- **Специализированная база данных**
(робот отсутствует)
- **Метасайт** -
небольшой по объему сборник ссылок на веб-страницы
- **Портал** -
многопрофильный сайт, предлагающий широкий спектр информационных услуг

Видимый веб:

30-45 млрд. страниц (2013 г.)

Скрытый веб (глубокий, невидимый, темный) -
онлайновые ресурсы, не попавшие в индексы
универсальных поисковых систем.

- Информация в базах данных
- Защищенная паролями и т.п.
- Запрещенная к индексированию владельцами
- Страницы, формируемые динамически
- Информация в нетекстовых файлах
- (Свежая, поэтому еще не проиндексированная)

*Скрытого в **сотни** раз больше, чем видимого*

Web 2.0, Web 3.0

- (Web 1.0) — условный термин;
“автор пишет, читатель читает”
- Web 2.0 — интерактивные сайты, где пользователи изменяют содержание; социальные сети; вики; блоги; онлайн-прикладные программы.
- Web 3.0 — предполагаемая следующая стадия развития, включающая «семантический веб»

Семантический веб будет основан на компьютеризованном распознавании **смысла** информации в документах.

Browse — Search

Два метода работы с онлайн-ресурсами:

- **Browse (перелистывание)** — движение по ссылкам.
- **Search (поиск)** — целенаправленное извлечение с помощью программы.

Браузер (browser) — прикладная программа, предназначенная для работы с веб-ресурсами.

*MS Internet Explorer (Обозреватель),
Mozilla **Firefox**, Opera, Google Chrome*

О достоверности информации

Традиционная vs. онлайн

- **Печатная литература**
автор известен
контроль со стороны издателя
- **Научная литература**
система **рецензирования** (peer review)
- **Веб-источники**
анонимность, отсутствие контроля – почти норма

Достоверность информации лежит в широких пределах:
от объективной - до субъективной,
от полностью достоверной - до ложной
и до намеренно сфальсифицированной

Оценка ресурса

В основе оценки онлайн-источника лежат известные критерии оценки печатных источников:

Репутация автора;

Контроль качества;

Объективность изложения;

Актуальность.

Плюс веб-специфика:

- *Рекламные* блоки могут казаться частью документа.
- *Отсканированный* и оптически распознанный текстовый материал редко выверяется корректорами.
- Содержание веб-страницы может быть изменено *несанкционированно* (атака хакера, прихоть администратора).
- Проблемы субъективности/достоверности особенно остро проявляются в *форумах* и *блогах*.

Стиль

Лингвистика

Явные признаки низкокачественного ресурса:

- Обилие опечаток и грамматических ошибок.
- Развязный стиль изложения.

Дизайн

Эксперт тщательно оценивает содержание, а обычный потребитель больше доверяет внешнему виду страницы.

Формальный анализ URL

Доменное имя

достоверность выше:

.gov .edu .ac.uk . ac.jp

достоверность ниже:

narod.ru

Папки

повысить бдительность:

~... private, members

Предпочтительны

Сайты:

- университетов,
- научных обществ,
- научных издательств,
- официальных патентных бюро,
- авторитетных коммерческих организаций,
- персональные сайты ученых.

Стремимся работать с **первоисточниками** и интенсивно используем **свой** мозг

Текстовые базы данных

- База данных (database) -
упорядоченный информационный массив,
состоящий из стандартных блоков.

Классификация по типу содержимого:

текстовые,
числовые,
формульные,
...

Структура базы данных (с точки зрения пользователя)

- **Запись (record)** - стандартный блок информации

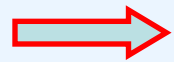
- **Поле (field)** - смысловой фрагмент записи

Поля:
текстовые,
числовые
и др.

Поле	Значение
Заглавие	Химия сегодня и завтра
Издано в	Мн. : Университетское, 1987.
Примечания	Библиогр.: с. 126-127 (42 назв.). 9630 экз.
Тематика	ХИМИЯ ХІМІЯ
УДК	54
ГРНТИ	31.01
Автор (лицо/организация)	Свиридов, Вадим Васильевич

Запись в каталоге библиотеки

- Поисковая программа
(search and retrieval software)



имеет страницу с **поисковым бланком**,
предназначенным для формулирования
запроса

- **Запрос (query)** -
поисковое задание, содержащее поисковые термины
и инструкцию по их интерпретации программой

Пример запроса:

натрий

Заполняем
поисковый бланк:

Поисковая программа ищет в своей базе данных те записи, в которых присутствует слово **натрий**

Список
обнаруженных
записей
выводится на
экран

[Натрий](#) - [[Translate this page](#)]

Натрий - жизненноважный межклеточный и внутриклеточный элемент, участвующий в ...
Потребность в **натрии** минимально составляет около 1 г/сут и в значительной ...
www.sunduk.ru/Encycl/ChemFood/C027.htm - [Cached](#) - [Similar](#)

[НАТРИЙ](#) - [[Translate this page](#)]

Натрий-22 с периодом полураспада 2,58 года используют в качестве источника позитронов. **Натрий-24** (его период полураспада около 15 часов) применяют в ...
www.krugosvet.ru/enc/nauka_i_tehnika/.../NATRI.html - [Cached](#) - [Similar](#)

Поиск - не по смыслу, а по факту наличия термина!

Синтаксис запроса в текстовых базах данных

Нет стандартного синтаксиса запроса.

У каждой программы **свои** правила.

Иногда правила совпадают
(но необязательно, что полностью).

Бывает, что некоторые элементы
разными поисковыми программами
воспринимаются *с точностью до наоборот*.

Логические (Булевы) операторы

- **AND**
натрий AND калий

& , ...

- **OR**
натрий OR калий

| , ...

- **NOT**
натрий NOT калий

- , (andnot, and not, but not)

варианты
обозначений

Оператор по умолчанию (default operator)

Пример: Обе записи равнозначны, если AND – по умолчанию:

натрий AND калий

натрий калий

Порядок выполнения операций

- Сначала: NOT и AND, затем: OR
- Если нужно, порядок меняют круглыми скобками

Пример:

Найти записи, в которых:
обязательно присутствует **натрий** или **калий** и
обязательно присутствует **фосфат** или **силикат**

Правильно:

(натрий OR калий) AND (фосфат OR силикат)

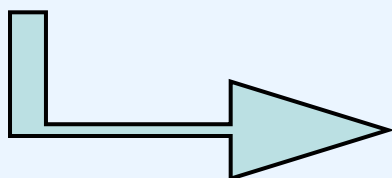
Неправильно:

натрий OR **калий AND фосфат** OR силикат

Операторы расстояния - 1

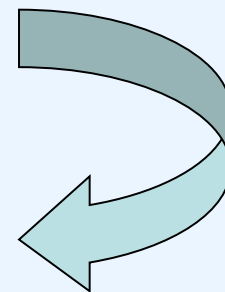
- Кавычки

Пример: "фосфат натрия"



два алгоритма:
фраза из 2 слов *или*
строка из 13 символов

"фосфат_натрия" \neq "фосфат__натрия"

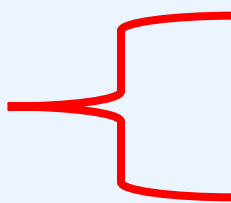
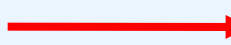


(символом подчеркивания обозначен пробел)

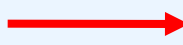
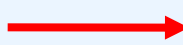
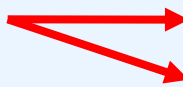
Операторы расстояния - 2

- WITH/n , NEAR/n (W/n, N/n, WITH, ...)

Пример: **aaa WITH/3 ббб**

		aaa ббб	(1)
извлекаются		aaa ввв ббб	(2)
		aaa ввв ггг ббб	(3)
не извлекаются			aaa ввв ггг ддд ббб

Пример: **aaa W/1 ббб**

извлекается		aaa ббб
не извлекается		ббб aaa
aaa N/1 ббб		
извлекаются		aaa ббб
		ббб aaa

Шаблон - 1

* ("звездочка")

заменяет **любое число** символов (в т. ч. нулевое)

Примеры: **фосфат***

фосфат, фосфатами, фосфатирование, ...

хлор*

хлор, хлорид, ...

НО: хлорофилл

***фосфат**

фосфат, **ди**фосфат, **поли**фосфат, ...

Wildcard. Truncation (right-hand, left-hand) - Усечение

Шаблон - 2

? (вопросительный знак), # (решетка)
заменяет **ОДИН** СИМВОЛ

Пример: **бут?н**
бут**а**н, бут**е**н, бут**и**н, бут**о**н

Как правило:

При шаблоне оставлять не менее трех букв.
Не использовать шаблон внутри кавычек.


Шаблон увеличивает количество
информационного мусора в результатах поиска

Stemming

- **Stemming** – режим работы поисковой программы, при котором происходит **учет грамматических форм** терминов (**учет морфологии, учет словоформ**)

Пример: **фосфат** 

фосфат, фосфатами, фосфатный, ... (полифосфат - ?)

Пример: **write** 

write, writes, writing, wrote

Не проводить stemming:

"фосфатами"

Стоп- слова

- **Стоп-слова (stopwords)** - слова, которые при поиске не учитываются.

Это слова, не несущие самостоятельной смысловой нагрузки, но особенно часто встречающиеся в тексте:
предлоги, союзы, артикли и т. п.

Пример:

~~**The Analyst**~~

Включить стоп-слово в поиск:

"The Analyst"

Регистр букв

- Абсолютное большинство поисковых программ нечувствительно к регистру букв – для них

строчные и заглавные буквы в запросе **равнозначны**.

Пример:

фосфат AND силикат

фосфат and силикат

фОсФаТ aNd СиЛиКаТ

годится любой вариант

Указание поля поиска

- Поиск можно сделать более эффективным, если проводить его не по записям в целом, а только по избранным полям.

Для этого в запросе рядом с поисковым термином указывают код соответствующего поля.

Коды полей в разных базах данных – разные.

Примеры:

ttl/фосфат

ttl/фосфат and натрий

фосфат filetype:pdf