

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**

10 недель:

Лекция (2 час.) +
+ **9** занятий x **4** час.

(минилекции + практикум + 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, 1999-2016. <http://www.abc.chemistry.bsu.by>



Презентации к лекциям

1 семестр (н.-пр., н.-пед., охр ОС, лек., фонд., ХВЭ)
Лекции: **1 2 3 4**

1 семестр (фарм.)
Лекции: **1 2 3**

2 семестр
Лекции: **1 2 3**



Практикум

1а. Текстовый поиск информации

1б. Патентные базы данных

2. Структурный поиск

Учебные программы

Справочные материалы



ABC Chemistry : бесплатные научные журналы по химии - каталог постоянно доступных полнотекстовых онлайн-журналов и информация о временно доступных журналах



Хімічны часопісы праз камп'ютарную сетку БДУ. Анлайн-базы дадзеных у бібліятэках Беларусі



25 стартовых точек поиска химической информации



ABC-Chemistry.org : Directory of Free Full-Text Journals in Chemistry - английская версия нашего каталога



СУПЕР ХІМІКІ

Суперхімікі : Хімічныя алімпіяды школьнікаў Беларусі : ілюстраваная гісторыя

Бюллетень химической информации

06.05.16. В Центральной научной библиотеке НАНБ открыт доступ к коллекции из 53 полнотекстовых российских журналов, расположенной в научной электронной библиотеке **eLIBRARY.RU**. Доступ к журналам осуществляется только с компьютеров библиотеки.

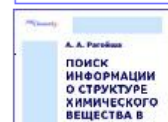
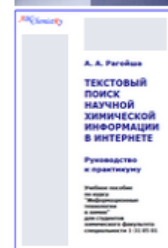
31.03.16. **100 статей, которые могут изменить мир**, опубликованные в 2015 году в журналах издательства *Springer*, - открыты до 15 июля 2016 г.

01.03.16. На неопределенный период времени открыты спецвыпуски журналов: **Bioorganic & Medicinal Chemistry** V.23 #11 (2015) (Emerging Approaches for the Design and Synthesis of Innovative Small Molecule Libraries), **Food Chemistry** V.193 (2016) (10th International Food Data Conference (IFDC): Joining nutrition, agriculture and food safety through food composition).

23.02.16. Журнал **Chemie in unserer Zeit** по случаю своего 50-летия открыл бесплатный доступ к подборке своих лучших публикаций **последних лет**. Кроме того, **№1 за 2016 г.** открыт как *Sample Issue*. Все вместе - это идеальный материал для сдачи "тысяч" по немецкому языку.

Google™ Custom Search

Search



Терминология

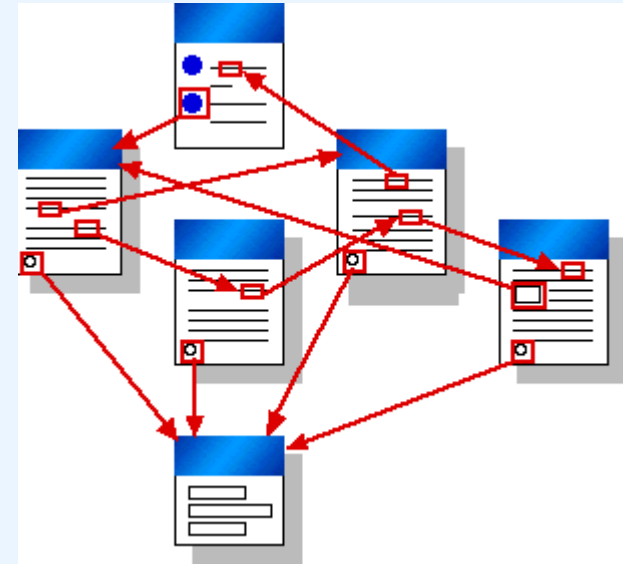
WWW

- **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
бел, рф
тут.бел = tut.by

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>

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.

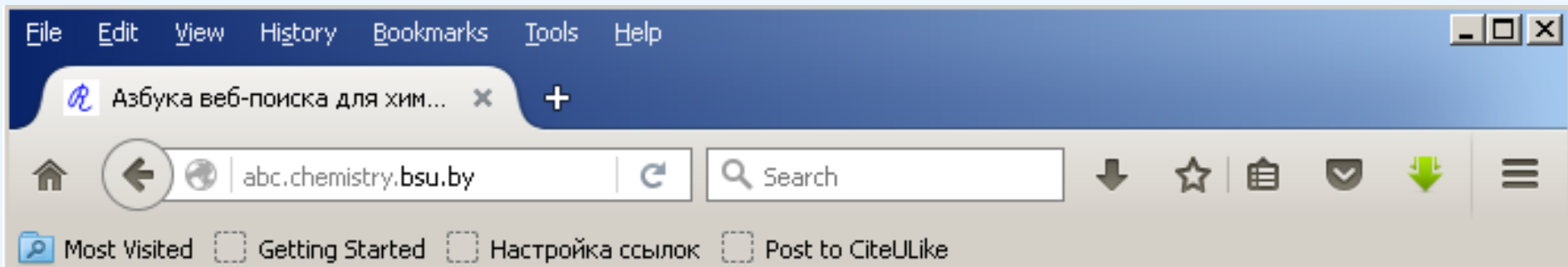
Browse — Search

Два метода работы с онлайн-ресурсами:

- Browse (перелистывание)
- Search (поиск)

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

Mozilla Firefox



Указатели ресурсов

Поисковая система

Search engine

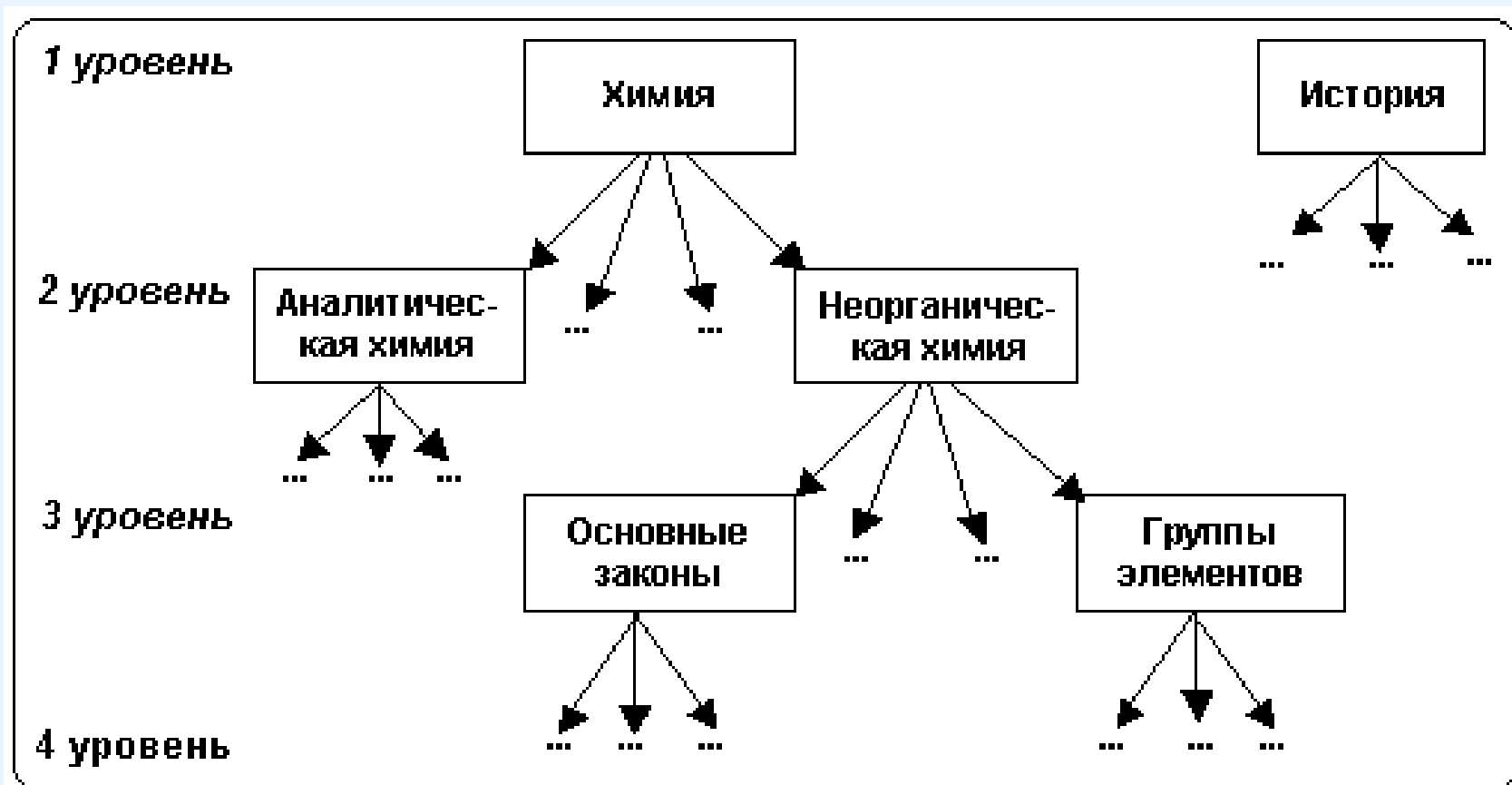
- робот (паук)
- индекс (база данных)
- поисковая программа, веб-интерфейс

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

Синтаксис запроса в текстовых базах 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

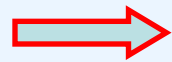
- **Поле (field)** - смысловой фрагмент записи

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!

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)

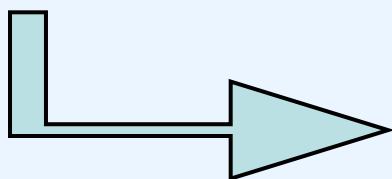
Неправильно:

натрий OR **калий AND фосфат** OR силикат

Операторы расстояния - 1

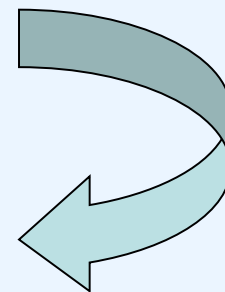
- Кавычки

Пример: "фосфат натрия"



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

"фосфат_натрия" \neq "фосфат__натрия"

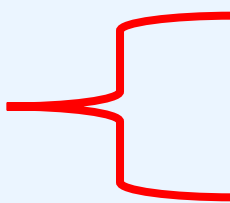
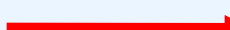


(символом подчеркивания обозначен пробел)

Операторы расстояния - 2

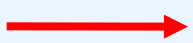
- WITH/n , NEAR/n (W/n, N/n, WITH, ...)

Пример: **aaa WITH/3 ббб**

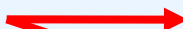
		aaa ббб	(1)
извлекаются		aaa ббб	(2)
		aaa ббб ббб	(3)
		aaa ббб гбб ббб	(4)
не извлекаются		aaa ббб гбб дбб ббб	(4)

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:

ttd/фосфат

ttd/фосфат and натрий

фосфат filetype:pdf

Поисковый бланк и список результатов поиска

Поисковый бланк - 1

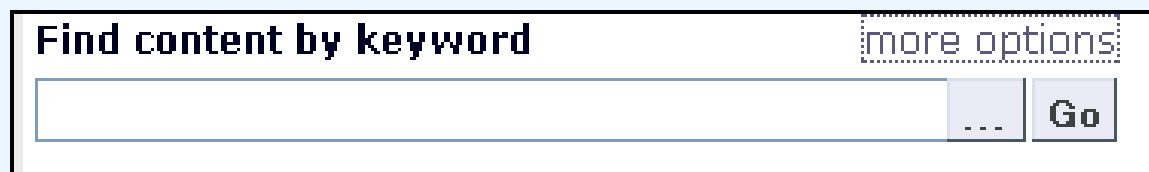
- Простейший, Basic, Quick, Simple
- Расширенный, Advanced, Expert

Классификация условна – в рамках определенной базы данных.

Обычно такие бланки называют **Quick Search**



A search form with two radio buttons. The first radio button is selected and labeled "All Content". The second radio button is unselected and labeled "Publication Titles". Below the radio buttons is a text input field. To the right of the input field is a "Go" button.



A search form titled "Find content by keyword". It features a text input field, a "more options" link, and a "Go" button.

Поисковый бланк - 2

Пример бланка

Quick Search:

Поиск по двум разным полям; использование булевых операторов; выбор временного интервала.

Query [\[Help\]](#)

Term 1: in Field 1:

Term 2: in Field 2:

Select years [\[Help\]](#)

Поисковый бланк - 3

Пример бланка

Advanced Search:

All Sources	Journals	Books	Advanced Search
Term(s):	<input type="text"/>	within:	<input type="text" value="Title"/>
AND	<input type="text"/>	within:	<input type="text" value="Authors"/>
Include:	<input checked="" type="checkbox"/> Journals	<input checked="" type="checkbox"/> All Books	
Source:	<input type="text" value="All sources"/>		
	Select one or more:		
Subject:	<input type="text" value="- All Sciences -"/> Agricultural and Biological Sciences Arts and Humanities Biochemistry, Genetics and Molecular Biology		Hold down the key) to select m
Dates:	<input checked="" type="radio"/> All Years	<input type="radio"/> 1999	to: <input type="text" value="Present"/>
Search Clear Recall Search			

Элементы бланка

The screenshot shows a search interface with the following elements:

- Navigation tabs: **All Sources** (highlighted), **Journals**, **Books**.
- Term(s):** An empty text input field.
- AND** (operator) with a dropdown arrow.
- Include:** Two checked checkboxes for **Journals** and **All Books**.
- Source:** A dropdown menu showing **All sources**.
- Subject:** A dropdown menu with the text "Select one or more:" above it. The selected item is **- All Sciences -**. Other visible items include **Agricultural and Biological Sci**, **Arts and Humanities**, and **Biochemistry, Genetics and M**.
- Dates:** Radio buttons for **All Years** (selected) and **1999** (with a dropdown arrow), followed by the text "to:".
- Buttons at the bottom: **Search**, **Clear**, and **Recall Sea**.

- Графа бланка (редактируемая графа, редактируемое поле).
- Список.
- Меню (выпадающий список).
- Переключатель.
- Выключатель.

- Текстовые пояснения.
- Ссылка на иной бланк.
- Ссылка на блок инструкций.
- Кнопка начала поиска.

Список результатов поиска

Пользователь получает результаты поиска в форме списка обнаруженных документов.

Список может быть сформирован:

- по алфавиту,
- в хронологическом порядке
(прямом или обратном),
- по релевантности.

Релевантность

Релевантность документа –

степень соответствия его поисковому заданию.

При расчете релевантности учитываются:

- количество поисковых терминов в документе,
- расстояние между ними в тексте,
- число упоминаний каждого из них,
- их плотность,
- их порядок расположения,
- их место – в начале записи или в конце,
- и др.

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:

livejournal.com

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