

# РОМИП в 2004 году



М.С. Агеев, НИВЦ МГУ  
Губин М.В., ИК «Кодекс»  
Добров Б.В., НИВЦ МГУ  
Кураленок И.Е., СПбГУ  
Некрестьянов И.С., СПбГУ  
Плешко В.В., Гарант-Парк-Интернет  
Сегалович И.В., Яндекс  
Шабанов В.И., Рамблер Интернет Холдинг

# Что такое РОМИП?

## Российский семинар по оценке методов информационного поиска

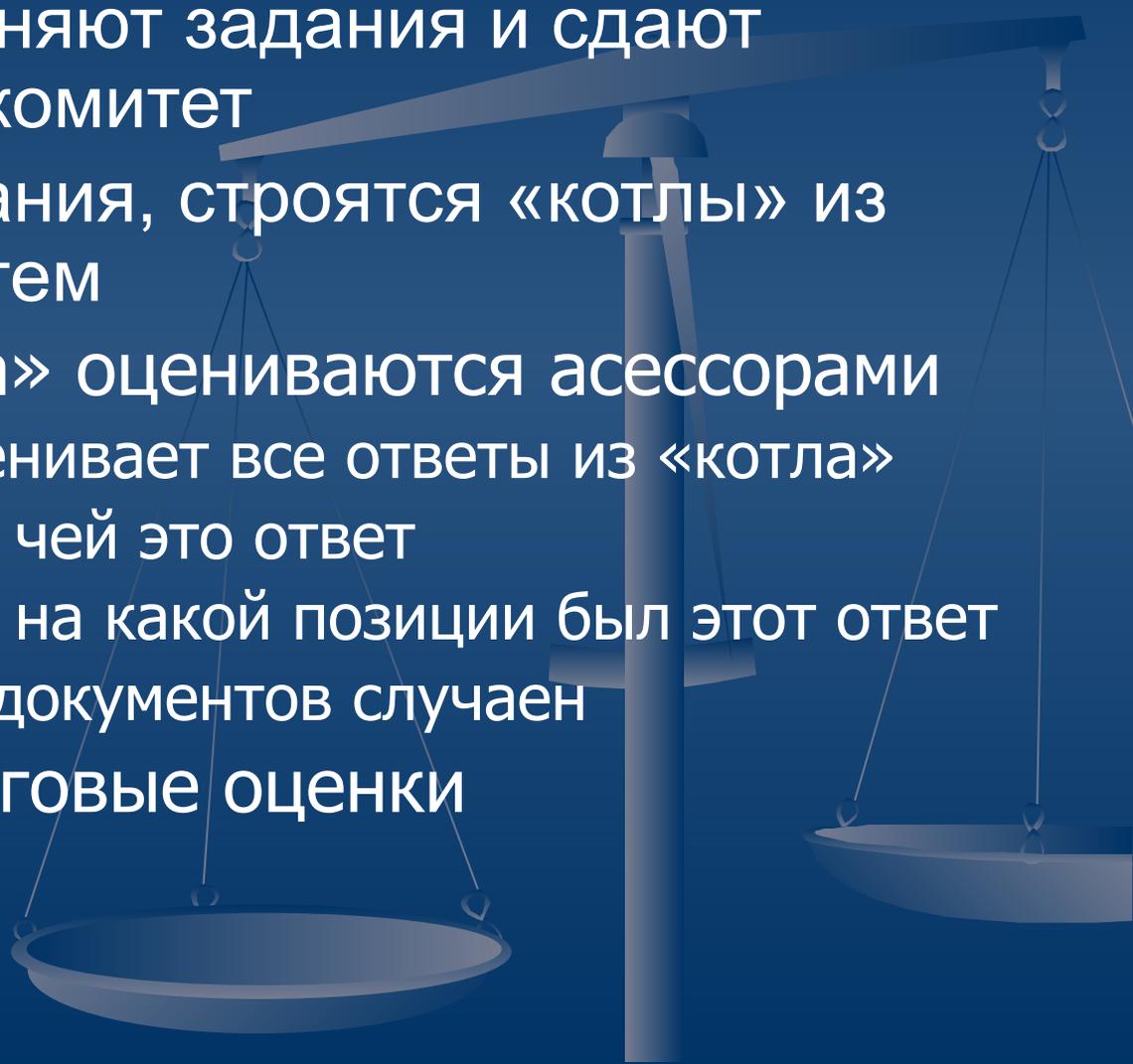
- Русскоязычные задания
- Крупные коллекции
- Использование апробированных подходов
- Равноправие и анонимность участников
- Использование независимых экспертов для оценки результатов поиска
- Возможность повторного использования

# Зачем это надо?

- создание публично доступных русскоязычных тестовых коллекций
- проведение независимой оценки методов информационного поиска, ориентированных на работу с русскоязычной информацией
- формирование среды для исследования проблем поиска:
  - «правил игры» - этических норм представления и использования результатов
  - требований к оформлению текстовых коллекций для тестирования

# Обобщенная процедура оценки

- Участники выполняют задания и сдают результаты в оргкомитет
- Для каждого задания, строятся «котлы» из ответов всех систем
- Ответы из «котла» оцениваются ассессорами
  - Один ассессор оценивает все ответы из «котла»
  - Ассессор НЕ знает чей это ответ
  - Ассессор НЕ знает на какой позиции был этот ответ
  - Порядок выдачи документов случаен
- Вычисляются итоговые оценки



# Эволюция

## 2003

- 1 коллекция (Веб)
- 2 дорожки
  - Поиск
  - Классификация
- Участники
  - 9 заявок
  - 7 дошло до финиша
  - 14 прогонов
- 550 часов оценки

## 2004

- 3 коллекции
- 5 дорожек
  - 3 новых
- Участники
  - 11 заявок
  - 9 дошло до финиша
  - 34 прогона
- 1300 часов оценки

# Коллекции

## Коллекция Narod.Ru

728 000 страниц,  
22 000 сайтов

Предоставлена «Яндекс»

## Коллекция Legal

61 000 HTML страниц,  
1.6 Гб

Предоставлена «Кодекс»

## Коллекция DMOZ

300 000 страниц  
(не более 500 страниц с одного сайта)  
Область применения: обучающее множество

# Задачи: поиск

## Задание:

Для каждого запроса вернуть упорядоченный список (до 100) документов

В «котлах» учитывались 50 первых ответов.

### Narod.ru

24250 запросов из журналов Яндекс и Рамблер.  
Оценка: 48 + 19 повторно.  
Альтернативная оценка с и без учета расширенных описаний.

### Legal

13000 запросов из журналов Кодекс и Парк.Ру .  
Классы запросов:  
50 “понятия” + 41 “документы”

# Задачи: классификация

## Задание:

Задан список категорий и обучающая выборка.  
Для каждого сайта/документа вернуть список до  
5 категорий к которым он относится.

### DMOZ/Narod.ru

Классификация *Веб-сайтов*  
247 категорий из каталога DMOZ

Оценка: 38 категорий

### Legal

163 категории  
13772 обучающих примера

Оценка: 12 вручную,  
40 «сравнение с эталоном»

# Задачи: поиск фактов

## Задание:

найти все события связанные с персоной.  
Ответ - фрагмент текста до 300 символов,  
описывающий это событие  
(текст + ссылка на его положение)

- Источник:  
«Кроссворд-кафе»

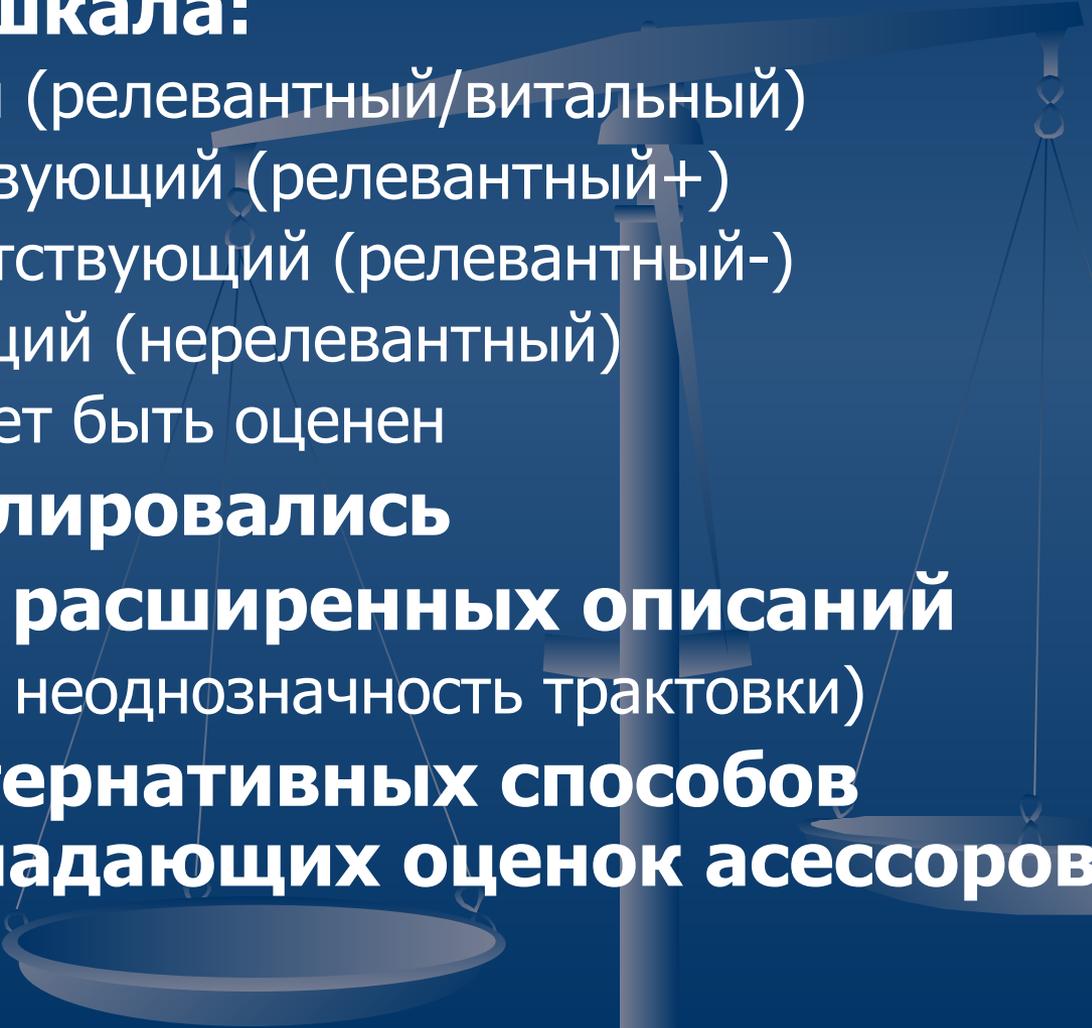
- 5052 задания

- Оценивалось: 109

- Попытка проверки границ выделения

```
<task id="qa1109">  
  <variant>Владимир Ильич Ленин</variant>  
  <variant>Владимир Ильич Ульянов</variant>  
  <description>  
    вождь мирового пролетариата  
  </description>  
</task>
```

# Оценка: детали



- **Многозначная шкала:**
  - Соответствующий (релевантный/витальный)
  - Скорее соответствующий (релевантный+)
  - Возможно соответствующий (релевантный-)
  - Не соответствующий (нерелевантный)
  - Документ не может быть оценен
- **Все оценки дублировались**
- **Использование расширенных описаний**  
(Цель – упразднить неоднозначность трактовки)
- **Несколько альтернативных способов слияния несовпадающих оценок ассессоров**

# Достижения РОМИП'2004

- Расширение количества участников
- Новые коллекции, а также наборы заданий и таблицы релевантности (доступны не только участникам РОМИП)
- Увеличение объема выполненных исследований повышение уровня результатов
- Отчеты участников по результатам выполненных исследований (доступны на [romip.narod.ru](http://romip.narod.ru))
- Соглашение об использовании коллекций
- Поддержка РФФИ

# РОМИП'2005

- Новая коллекция: новостная  
(предоставлена Яндекс)  
(24000 сообщений от 16 агентств за 3 периода)
  - 6 новых дорожек
    - Контекстно-зависимое аннотирование
    - Фактографический поиск
    - Структурирование новостного потока
    - Классификация Веб-страниц
  - 45+ заявок от 15 участников
  - Статус:
    - сбор результатов от участников
    - очная часть: ~ 1 октября
- 