Проект общедоступного многопользовательского корпуса

Проблемы и перспективы

Дмитрий Грановский

mathlingvo

Зачем ещё один?

У существующих корпусов:

- Авторские права
  - На текст
  - На разметку
- **Административные ограничения**
- Недоступна сама разметка

Что будет уметь?

- Просмотр и редактирование разметки (многопользовательский)
- Возможность скачивания в разных форматах
- Встроенный поиск по популярным запросам
- Обработка «тяжёлых» запросов

mathlingvo

Кто будет делать?

- Техническая часть – группа разработчиков

- Наполнение – **пользователи**

Откуда брать тексты?

- Тексты с истекшим авторским правом
- Тексты, на которые авторское право не распространяется
- Тексты под свободной лицензией (e.g. Википедия)

Остальное – по договорённости с правообладателем

Задачи инструментария

1. Редактирование
2. Хранение
3. Поиск

Как можно использовать?

- Статистические исследования
- Машинное обучение (как обучающий корпус)
- Тестовый корпус для других ресурсов

Проблемы

- Почему эта модель доступа должна работать?
- Как обеспечить совместную работу многих людей? (*многопользовательский*)
- Как обеспечить целостность данных?

Проблемы

- Откуда взять столько квалифицированных редакторов?
- Как бороться с ошибками редактирования?
- Как не делать разметку с нуля?
- Как унифицировать разметку?

Ваши вопросы

СПАСИБО!

Дмитрий Грановский

d-granovsky@yandex.ru

mathlingvo

http://mathlingvo.ru