

## Лекция 4. Поиск и исправления ошибок с помощью OpenRefine

**Наталья Иванова**

Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН

**BioDATA**



## Что такое OpenRefine?

*“Мощный инструмент для работы с “сырыми” данными”*

Возможность быстро получить представление о больших массивах информации

- Фильтровать и объединять данные
- Преобразовывать данные в нужный формат, делать базовые расчеты
- Находить ошибки и неожиданности — например, слишком большие цифры, слова вместо чисел, пустые значения
- Автоматически находить потенциальные опечатки и несоответствия в названиях, позволяя приводить записи к единому виду (кластеризация текстовых записей)



## Что такое OpenRefine?

*“Мощный инструмент для работы с “сырыми” данными”*

- Программа открывается с помощью вашего интернет-браузера, но при этом является автономным приложением, поэтому подключение к Интернет не требуется.
- OpenRefine совместима с Windows, Mac и Linux.

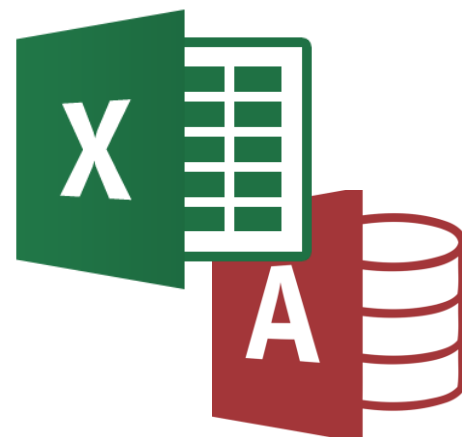


## Почему не?

База данных



Excel-подобная практика



## Сравнение функций

### База данных



- Необходима инфраструктура и навыки программирования для редактирования
- Нет простой визуализации

### Excel



- Обычно редактируется одна ячейка
- Полезен для документирования данных и выполнения операций
- Данные не всегда видны
- Недостаточная визуализация

### OpenRefine



- Одновременное редактирование многих ячеек
- Простой поиск и преобразование
- Интерактивная визуализация

## Преимущества

### OpenRefine



- Редактирование многих ячеек одновременно
- Простой поиск и преобразование
- Интерактивная визуализация
- Бесплатно
- Настольное приложение (работает оффлайн)
- Можно отменить любое действие / преобразование
- Использование API
- Экспорт и импорт файлов нескольких типов
- Сводки / фильтры
- Использование регулярных выражений
- Расширения
- Большое сообщество разработчиков (расширения, руководства, и др.)

## Полезные ссылки

<https://github.com/OpenRefine>

Руководство по проверке видовых названий в OpenRefine:

[https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr\\_JEGTvhRo/edit](https://docs.google.com/document/d/1tkDRXIYhmassYAk5T4v5oac5prF0jAiSMr_JEGTvhRo/edit)

-  
Руководство по работе с высшей таксономией в OpenRefine:

[https://docs.google.com/document/d/1XZ\\_pM9gldQzHzl8wfUCVea-52yub5T\\_3tc-snBgPRa0/edit](https://docs.google.com/document/d/1XZ_pM9gldQzHzl8wfUCVea-52yub5T_3tc-snBgPRa0/edit)

-  
OpenRefine – документация для пользователей:

<https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users>

-  
Большой список дополнительных ресурсов по OpenRefine:


<https://github.com/OpenRefine/OpenRefine/wiki/External-Resource>

# Практическое занятие

ВЕРИФИКАЦИЯ, ИСПРАВЛЕНИЕ  
И СТАНДАРТИЗАЦИЯ ДАННЫХ



## Условные обозначения

- Формулы (скопировать-вставить) `Cell.recon.match.id`
- Команды в OpenRefine `Edit column`
- Названия столбцов `nameRecon`
- Полезные ссылки [www.gbif.org](http://www.gbif.org)
- Меню столбца 

# Упражнения

- Загрузка файлов в OpenRefine
- Сводка и массовое редактирование: быстрое исправление опечаток и удаление лишних пробелов, поиск дубликатных записей
- Применение фильтров: поиск опечаток и проверка регистра в таксономических данных
- Кластеризация: поиск ошибок в написании регионов
- Сохранение результатов

## Лекция 4. Поиск и исправления ошибок с помощью OpenRefine

**Наталья Иванова**

Институт математических проблем биологии РАН – филиал ИПМ им. М.В. Келдыша РАН

**BioDATA**

