

База данных vs. таблички, когда и что выбирать?

Г. Мороз

2 октября 2022

Четвертый выездной семинар Оплинга



Постановка проблемы

Мы все чаще работаем с разнородными данными, которые нужно:

- анализировать (иногда повторно)
- хранить
- пополнять
- исправлять
- публиковать
- цитировать

Сегодня мы будем делать вид, что это можно делать при помощи

- обычных таблиц Excel/Libre Office

Постановка проблемы

Мы все чаще работаем с разнородными данными, которые нужно:

- анализировать (иногда повторно)
- хранить
- пополнять
- исправлять
- публиковать
- цитировать

Сегодня мы будем делать вид, что это можно делать при помощи

- обычных таблиц Excel/Libre Office
- баз данных

Постановка проблемы

Мы все чаще работаем с разнородными данными, которые нужно:

- анализировать (иногда повторно)
- хранить
- пополнять
- исправлять
- публиковать
- цитировать

Сегодня мы будем делать вид, что это можно делать при помощи

- обычных таблиц Excel/Libre Office
- баз данных

Постановка проблемы

Мы все чаще работаем с разнородными данными, которые нужно:

- анализировать (иногда повторно)
- хранить
- пополнять
- исправлять
- публиковать
- цитировать

Сегодня мы будем делать вид, что это можно делать при помощи

- обычных таблиц Excel/Libre Office
- баз данных

Фрэнсис М. Тайерс: *“Вроде работает. Не тормозит. Пока можешь, лучше занимайся лингвистикой.”*

Проще всего про это думать как о данных, у которых есть некоторая схема, позволяющая делать что-то с данными в соответствии с этой схемой. Выделяют разные типы баз данных:

- иерархическая (например, отношения язык - словарь - лемма - значение)
- сетевые (например, перекрестные ссылки в научных статьях)
- и другие

Часто базы данных требуют какого-нибудь отдельного языка запросов: SQL или какого-нибудь языка для анализа данных (Python, R).

Чем хороши базы данных для науки?

- часто данные хранятся в оптимальном формате без лишних повторений

Чем хороши базы данных для науки?

- часто данные хранятся в оптимальном формате без лишних повторений

author	title	year	character
Н. В. Гоголь	Мертвые души	1842	Чичиков
Н. В. Гоголь	Мертвые души	1842	Сабакевич
Н. В. Гоголь	Мертвые души	1842	Плюшкин
...
А. С. Пушкин	Евгений Онегин	1842	Онегин
А. С. Пушкин	Евгений Онегин	1842	Ленский
...

vs

Table 1

title	character
Мертвые души	Чичиков
Мертвые души	Сабакевич
Мертвые души	Плюшкин
...	...
Евгений Онегин	Онегин
Евгений Онегин	Ленский
...	...

Table 2

author	title	year
Н. В. Гоголь	Мертвые души	1842
А. С. Пушкин	Евгений Онегин	1842

Чем хороши базы данных для науки?

- часто данные хранятся в оптимальном формате без лишних повторений
- не делает орфографических/топографических ошибок
 - латиница вместо кириллицы
 - автозамена в Excel на большие буквы, даты и т. п.
- часто к базам данных можно прикрутить версионирование, так что каждое изменение сохраняется
- часто к базам данных легко сделать визуальный интерфейс
- часто к базам данных можно устроить совместный доступ пользователей с разными правами (чтение, редактирование и т. п.)

Чем хороши базы данных для науки?

- часто данные хранятся в оптимальном формате без лишних повторений
- не делает орфографических/топографических ошибок
 - латиница вместо кириллицы
 - автозамена в Excel на большие буквы, даты и т. п.
- часто к базам данных можно прикрутить версионирование, так что каждое изменение сохраняется
- часто к базам данных легко сделать визуальный интерфейс
- часто к базам данных можно устроить совместный доступ пользователей с разными правами (чтение, редактирование и т. п.)

ОДНАКО я считаю, что в большинстве случаев создание баз данных в лингвистике не стоит трудов и можно довольствоваться обычной табличкой.

Спасибо за внимание!