Как НКО выстроить процесс сбора и хранения данных

Цикл «Исследования НКО»
https://ngo-research.ru

Информационная культура

ФОНД ПРЕЗИДЕНТСКИХ ГРАНТОВ

В чем польза

Экономия ресурсов

- не надо получать специальные навыки или обращаться к техническим специалистам
- целенаправленная подготовительная и сопроводительная работа позволяет избегать лишних временных затрат и грубых ошибок

Преимущества

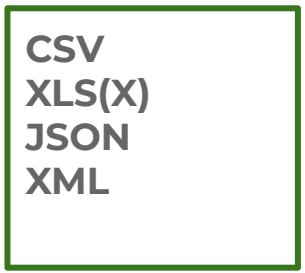
- порядок => не надо долго искать, переделывать, вспоминать
- возможность возвращаться к данным и использовать их повторно
- возможность публиковать данные
- возможность создавать на них полезные продукты

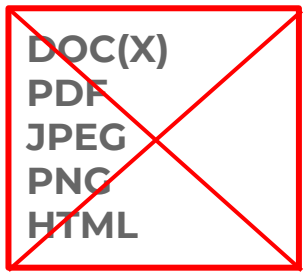
Хранение vs. представление

Аналогия: архив vs. музей

Хранение: упорядоченность, механизмы поиска, сохранность в первозданном виде

Представление: показ, встраивание в контекст, подчеркивание определенных аспектов





Шаги

- Определение структуры
- Выбор формата хранения и соответствующего ПО
- Создание интерфейса для сбора данных
- Сбор данных
- Соблюдение мер предосторожности для сохранения данных
- Описание данных

Структура

Выявление признаков описания однородных объектов предметной области.

Пример города России:

- наименование города
- наименование субъекта федерации
- код субъекта федерации
- численность населения
- признак города федерального значения
- официальный сайт администрации города
- дата актуальности

Структура

Демонстрация

Структура: данные в таблице

Одномерная структура: 1 строка - 1 объект

Пример: https://bit.ly/3bh4pJ7

	A	B	C	D	E	F	G
1	**Название города**	**Код субъекта**	**Наименование субъекта**	**Численность населения (человек)**	**Город федерального значения**	**Официальный сайт администрации**	**Год актуальности**
2	Санкт-Петербург	78	Санкт-Петербург	5398064	ИСТИНА	https://www.gov.spb.ru/	2020
3	Северодвинск	29	Архангельская область	181990	ЛОЖЬ	http://severodvinsk.info/	2020
4	Выборг	47	Ленинградская область	75355	ЛОЖЬ	http://www.city.vbg.ru/	2020
5	Архангельск	29	Архангельская область	346979	ЛОЖЬ	https://www.arhcity.ru/	2020
6	Москва	77	Москва	12678079	ИСТИНА	https://www.mos.ru/	2020
7	Санкт-Петербург	78	Санкт-Петербург	5383890	ИСТИНА	https://www.gov.spb.ru/	2019
8	Северодвинск	29	Архангельская область	182291	ЛОЖЬ	http://severodvinsk.info/	2019
9	Выборг	47	Ленинградская область	76389	ЛОЖЬ	http://www.city.vbg.ru/	2019
10	Архангельск	29	Архангельская область	348343	ЛОЖЬ	https://www.arhcity.ru/	2019
11	Москва	77	Москва	12615279	ИСТИНА	https://www.mos.ru/	2019

Структура: двухмерная таблица

	2020	2019
Санкт-Петербург	5398064	5383890
Северодвинск	181990	182291
Выборг	75355	76389
Архангельск	346979	348343
Москва	12678079	12615279

Плюсы: человекочитаемость, зрительная компактность
Минусы: неудобство и потенциальная некорректность машинной обработки

=> **такой формат хорош для представления, но не подходит для хранения данных**

	2019		2020		Субъект РФ
	Численность населения	Город федерального значения	Численность населения	Город федерального значения	
Санкт-Петербург	5383890	Да	5398064	Да	Санкт-Петербург
Северодвинск	182291	Нет	181990	Нет	Архангельская область
Архангельск	348343	Нет	346979	Нет	
Выборг	76389	Нет	75355	Нет	Ленинградская область
Москва	12615279	Да	12678079	Да	Москва
Итого	18606192	-	18680467	-	-

Структура: важность определений

Необходимо точное определение сущностей

Примеры:

Что имеется в виду под “городом”? Населенный пункт, имеющий статус города? Городская агломерация?

Что описывает один объект (одна строка таблицы)? Город вообще? Состояние города, актуальное на определенный год?

Структура: иерархическое расширение

Ситуация:

У города несколько вокзалов, и у каждого вокзала могут быть свои характеристики (название, год открытия, оператор...).

Решения:

- специальные иерархические форматы (XML, JSON)
- связанные таблицы

Структура: иерархическое расширение

Пример: JSON

```
{
	"town_name": "Санкт-Петербург",
	"region_code": "78",
	"region_name": "Санкт-Петербург",
	"population": 5398064,
	"is_federal": true,
	"administration_website": "https://www.gov.spb.ru/",
	"year": 2020,
	"rw_stations": [
		{
			"rw_name": "Финляндский вокзал",
			"rw_operator": "РЖД",
			"begin_year": 1870,
		},
		{
			"rw_name": "Московский вокзал",
			"rw_operator": "РЖД",
			"begin_year": 1847,
		},
	]
}
```

Структура: иерархическое расширение

Распределение по разным таблицам с возможностью связи

Название города	Код субъекта	Наименование субъекта	Численность населения (человек)
Санкт-Петербург	78	Санкт-Петербург	5398064
Северодвинск	29	Архангельская область	181990
Выборг	47	Ленинградская область	75355
Архангельск	29	Архангельская область	346979
Москва	77	Москва	12678079
Санкт-Петербург	78	Санкт-Петербург	5383890
Северодвинск	29	Архангельская область	182291
Выборг	47	Ленинградская область	76389
Архангельск	29	Архангельская область	348343
Москва	77	Москва	12615279

Город	Название вокзала	Операто
Санкт-Петербург	Балтийский вокзал	
Санкт-Петербург	Витебский вокзал	
Санкт-Петербург	Ладожский вокзал	
Санкт-Петербург	Московский вокзал	РЖД
Санкт-Петербург	Финляндский вокзал	РЖД
Северодвинск	ЖД вокзал Северодвинска	РЖД
Выборг	ЖД вокзал Выборга	
Архангельск	ЖД вокзал Архангельска	
Москва	Белорусский вокзал	
Москва	Казанский вокзал	
Москва	Киевский вокзал	
Москва	Курский вокзал	
Москва	Ленинградский вокзал	
Москва	Павелецкий вокзал	
Москва	Рижский вокзал	
Москва	Савёловский вокзал	
Москва	Ярославский вокзал	

Структура: желательно включить

- потенциальные связи с другими наборами данных
- связи между разными таблицами, если они дополняют друг друга

Табличные форматы: CSV vs. XLS(X)

CSV	XLS(X)
Полная прозрачность структуры и содержания	Много возможностей исказить структуру и содержание (в том числе непреднамеренно)
Довольно тяжелый	Относительно легкий в варианте XLSX
Можно прочитать с помощью самого разного ПО	Невозможно прочитать без специального ПО
Требует особых настроек при загрузке (кодировка, разделители)	Не требует специальных настроек при загрузке
Легко преобразуется в другие форматы	Возможность преобразования очень зависит от структуры конкретной таблицы

Табличный формат: программное обеспечение

LibreOffice (OpenOffice) Calc	• Бесплатный • Много инструментов для работы с таблицами • Довольно медленный
Таблицы Google	• Бесплатные • Довольно много инструментов • Допускают совместную работу • Хранятся онлайн • Больше уязвимы к взломам
Microsoft Excel	• Платный • Много инструментов для работы с таблицами

А также: Airtable (https://airtable.com/)

Сбор данных

Сбор готовых данных

- скачивание опубликованных наборов
- запрос данных
- получение данных с HTML-страниц

Сбор данных от людей

- опросы
- ручной сбор данных из разных неструктурированных источников

Сбор данных: HTML-таблицы

Демонстрация

Формулы из демонстрации:

- Импорт таблицы с кодами регионов из Википедии

```
=IMPORTHTML("https://ru.wikipedia.org/wiki/Коды_субъектов_Российской_Федерации";"table";1)
```

- Соединение текста из разных ячеек (конкатенация)

```
=concat(I2;" | ")
=concat(J2;H2)
```

- Разбивание на колонки строки с разделителем “|” с сохранением строкового формата цифр(например: “01 | Республика Адыгея”)

```
=split(REGEXREPLACE(A2;"^";"'");" | ";false)
```

Сбор данных вручную

Лучше всего вносить данные через какой-нибудь интерфейс.

Преимущества:

- не нужно лишний раз трогать таблицу, рискуя случайно изменить в ней значения
- ограничение типов значений, проверка
- возможность регулировать обязательность заполнения
- возможность задавать выбор значений из списка (=> корректность и сокращение времени на заполнение)

Формы

Самый распространенный и проработанный инструмент сбора данных.

Существует много вариантов, в том числе:

- Google-формы
- SurveyMonkey (https://www.surveymonkey.com/)
- Qualtrics (https://www.qualtrics.com/free-account/)

Формы

Демонстрация

Форма: ограничения

- Обязательность / необязательность ввода ответа
- Проверка типа значения (текст, число)
- Возможность выбора из заданных значений
- Возможность добавить “Другое”

Персональные данные

- Закон о персональных данных (152-ФЗ) (http://www.consultant.ru/document/cons_doc_LAW_61801/)
- Лучше не собирать
- Компромиссный вариант - собирать в добровольном порядке
- Если собирать, то важно проконсультироваться с юристами о формулировки условий конфиденциальности
- Анонимизация данных

Персональные данные: анонимизация

- Убрать всё, что позволяет идентифицировать конкретного человека
- Сохранить возможность идентифицировать, уникальность респондента

Имя донора	Email	Тип платежа
[illegible]	[illegible]@y[illegible]	разовое
[illegible]	[illegible]r@[illegible]	разовое
[illegible]	[illegible]@gr[illegible]	разовое
[illegible]	[illegible]@[illegible]	разовое
[illegible]	[illegible]@li[illegible]	разовое
[illegible]	[illegible]@[illegible]	разовое
[illegible]	[illegible]@li[illegible]	разовое
[illegible]	[illegible]@lis[illegible]	разовое
[illegible]	[illegible]@lis[illegible]	разовое
[illegible]	[illegible]v@gr[illegible]	разовое

ID	Тип платежа	Способ платежа
1	разовое	Банковская карта/регулярный платеж (CloudPayments)
2	разовое	Банковская карта (Яндекс.Деньги)
3	разовое	Банковская карта (Яндекс.Деньги)
4	разовое	Яндекс.Деньги
4	разовое	Банковская карта (Яндекс.Деньги)
4	разовое	Яндекс.Деньги
4	разовое	Яндекс.Деньги
4	разовое	Яндекс.Деньги
4	разовое	Яндекс.Деньги
5	разовое	Яндекс.Деньги

Хранение данных: как не испортить

- Работать только с копиями (не с оригиналом)
- Делать резервные копии
- Корректно загружать (особенно CSV: важность типов значений, кодировки, разделителей)
- Правильно сортировать

Загрузка CSV

Демонстрация

Метаданные и описание

- Делать расшифровки названий полей (еще на этапе создания структуры)
- Фиксировать, когда собраны данные
- Фиксировать срок актуальности
- Указывать методику сбора данных
- Хранить в предсказуемом месте

Вопросы?