# TEXTY.ORG.UA

# Відкриті дані: формати і правила створення

Київ, 2017

# Відкриті дані: формати і правила створення

## Робота з даними: А, Б, В

Щоб робота з даними була зручною для Вас, Ваших колег та всіх користувачів даних, потрібно дотримуватись кількох правил. Це заощадить Ваш час і збереже нерви.

### Як записувати дані у таблицю

Таблиця — це впорядкована сукупність стовпчиків та рядків.

Один рядок таблиці — це одна одиниця Ваших даних: мовою статистики — одне спостереження.

Один стовпчик — це одна змінна, тобто значення, яке змінюється від рядка до рядка.

Тобто, таблиця з об'єднаними комірками — це не дані. Принаймні, не відкриті дані. В такій таблиці неможливо відфільтрувати дані або здійснити математичні операції у стовпчиках.

Так само не є даними колір комірки чи шрифт у ній. Ці функції форматування можуть бути корисними для представлення даних, але незрозумілі для машини. Тому якщо інформація, закодована за допомогою кольору чи стиля шрифту, важлива, обов'язково додавайте її окремим стовпчиком.

		Чисельність безро в середньому за	бітного н І півріччя	аселення а 2017р.
	усьо	го, у віці 15-70 років	з ньог	о працездатного віку
	тис. осіб	у % до економічно активного населення відповідного віку	тис. осіб	у % до економічно активного населення відповідного віку
Україна	1709,7	9,6	1709,4	10,0
Вінницька	79,1	10,9	79,1	11,3
Волинська	53,0	12,7	53,0	12,8
Дніпропетровська	128,0	8,4	128,0	8,7

Приклад, в якому структура рядків і колонок порушена<sup>1</sup>:

Область	тис. осіб	%	Вік
Україна	1709,7	9,6	15-70 років
Україна	1709,4	10	працездатного віку
Вінницька	79,1	10,9	15-70 років
Вінницька	79,1	11,3	працездатного віку
Волинська	53	12,7	15-70 років
Волинська	53	12,8	працездатного віку
Дніпропетровська	128	8,4	15-70 років
Дніпропетровська	128	8,7	працездатного віку

Ті ж самі дані в простій табличній структурі:

Один рядок — це чисельність та частка безробітного населення певного віку в певній адміністративній одиниці. Застосувавши простий фільтр (в Екселі або іншій програмі) можна отримати значення рівня безробіття, наприклад, в Дніпропетровській області серед населення працездатного віку. Таблицю можна сортувати, здійснювати підрахунки й аналіз. Структуровані таблиці можна зіставляти та поєднувати одну з одною, використовувати значення змінних, як ключі для об'єднання таблиць, легко змінювати структуру таблиці для різних потреб. Тобто, правильно структурована таблиця придатна для аналізу даних.

З даних в такій формі легко отримати вибірку або зведену таблицю (pivot table) з сумарними значеннями, проте зворотній процес часто неможливий.

Варто пам'ятати, що люди і машини сприймають таблиці по-різному. Дані для презентації або звіту цілком можуть мати об'єднані комірки і порушувати табличну структуру, або містити порожні рядки для форматування. Проте дані для подальшого аналізу мають бути, насамперед, легкими для сприймання комп'ютером і роботи з ними.

При цьому, дані в кожному окремому стовпці таблиці обов'язково мають бути одного типу.

Одну і ту саму дату можна записати в декількох форматах: 31.12.2017, 2017-12-31, 31/12/2017 тощо. В одній колонці використовуйте лише один варіант написання, тоді в будь-якій програмі буде легко прочитати ці дані саме як дату. В роботі з географічними даними подібність теж важлива. «Вул.Т. Шевченка, буд.15, кв.15, Київ, 01001», та «01001, Шевченка 15-15» сприймаються комп'ютером як різні дані, хоч і позначають одне й те саме. Якщо знадобиться виділити з адреси лише місто, або лише номер будинку, при однаковому написанні адрес задачу легко виконати. Якщо вагаєтесь, як варто записати адресу, краще зробіть окремі змінні-стовпці для кожної її частини: назви населеного пункту, поштового індексу, назви вулиці, номеру будинка та інших. Адже об'єднати дані простіше, аніж розділити.

Отже, все, що можна записати однаково, варто писати однаково. Імена, назви, номери телефонів, числа (десятковий розділювач має бути однаковий), дати, час, адреси. Все, що тільки можливо, зберігайте в уніфікованому форматі.

Отже, правила гарного тону при роботі з таблицями й даними:

- В жодному разі не використовуємо об'єднані комірки.
- Один рядок даних відповідає одному спостереженню.
- Один стовпчик в таблиці одна змінна. Поширена помилка використовувати значення змінної (назви областей або групи за доходами) як стовпці даних таблиці.
- Використовуємо перенос рядка в текстовій комірці виключно у випадках, коли це необхідно.
- Звертаємо увагу та вірно обираємо формат комірок (текст, число, дата).
- Не поєднуємо кілька одиниць даних в одній комірці.
- Не використовуємо латинські літери замість кириличних, і навпаки. Наприклад, рядки «МАМА» (українськими літерами) та «МАМА» (латинськими) виглядають однаково, проте вони нетотожні для комп'ютера. Це особливо стосується великих літер при написанні ініціалів. Або, наприклад, неможливо відрізнити «MCXII» (українською) та «MCXII» (латиницею) – а для комп'ютера перше буде незрозумілою послідовністю літер, а друге – римським числом.
- На початку і в кінці комірки з даними не має бути пробілів. Самі пробіли не можуть повторюватись. Отже, два пробіли підряд — це помилка, а використовувати пробіли для відступу в абзацах чи вирівнювання комірок вкрай неправильно. Для очищення даних від зайвих пробілів використовуємо функцію СЖПРОБЕЛЫ (TRIM) в Excel. В інших програмах є аналогічні функції.
- Перевіряємо відсутність недрукованих символів в значеннях за допомогою функції ПЕЧСИМВ (CLEAN).
- Використовуємо типографські лапки (французькі лапки, так звані «ялинки») «», але якщо потрібно узяти в лапки фразу всередині інших лапок, використовуємо німецькі лапки ("німецькі лапки"). Наприклад: ТОВ «ВКФ "Промспецбуд"».
- Послідовно використовуємо знак апострофу 'замість інших варіантів ('``).

- Ставимо пробіли після розділових знаків, але не ставимо пробіли до розділових знаків. Вірно: Іванов І. І. Невірно: Іванов І. І., Іванов ІІ.
- Записуйте дати і час однаково. Бажано використовувати формат ISO8601, а саме [рік]-[місяць]-[день]. Наприклад, 2017-12-31.
- Використовуйте простий формат для чисел не слід вдаватися до зайвого форматування. І використовуйте крапку як десятковий розділювач: 1.99 замість 1,99. Якщо необхідно зберегти провідні нулі, форматуйте комірки з числами, як текст.
- Не треба писати текстові значення чи назви колонок ВЕЛИКИМИ ЛІТЕ-РАМИ (окрім абревіатур). В кожній програмі для роботи з таблицями є можливість зробити всі літери маленькими.
- Не залишайте порожніх комірок внесіть в них запис NA (not available).
   Обов'язково вкажіть причини виникнення таких значень в таблиці.
- Намагайтеся уявити, як саме будуть використовувати ваші дані, та як їх використали б саме Ви!
- Зробіть все, щоб користуватися даними було зручно та приємно!

#### Як почистити таблицю

В більшості випадків ідеальні дані, які задовольняють наші потреби, є результатом нашої ж кропіткої праці. Тож вміння чистити дані надзвичайно важливе. Без нього Ви просто не отримаєте даних, з якими можна працювати.

Перетворення «брудної» таблиці на чисті дані можна розділити на дві частини: 1) зміна структури даних; 2) виправлення помилок в значеннях комірок.

Ми перетворили структуру таблиці в прикладі з даними Держстату. Прості зміни структури можна робити в табличних редакторах, таких як Excel aбo Libre Office Calc. Далі ми розберемо основні функції на прикладах.

Маємо таблицю, де кілька змінних записано в 1 стовпчику. Це незручно і суперечить правилам роботи з таблицями (а вони існують, щоб полегшити всім життя). Розглянемо, що можна зробити за допомогою MS Excel; в інших табличних редакторах є схожі функції.

 А
 В

 1
 Назва і ЄДРПОУ підприємства
 Прибуток

 2
 КП "Щось і Дещо" 656565
 10000

 3
 ТОВ "Підприємство" 232323
 1 000 000

 4
 КП "Супер-комунальники" 555555
 65 тис.

Так виглядає таблиця з вигаданими брудними даними:

Бачимо, що код ЄДРПОУ знаходиться в одній колонці з назвою. Треба це змінити!

Насамперед ми маємо визначити, що розділяє ці два значення. Назва кожного підприємства закінчується лапками, за ними іде пробіл. Отже, використаємо ці знаки в якості розділювача стовпців.

Коли ми поділимо колонку на дві частини, розділювач зникне. Тож заздалегідь подбаємо про те, аби зберегти всі лапки в назві підприємства, замінивши «"» на «";»

А) Виділіть колонку, в якій хочете замінити одні значення на інші

Б) Натисніть Ctrl+H (або Правка — Знайти і замінити)

В) Введіть що заміняти — («"»), і на що — («";»).

Г) Натисніть «Замінити всюди». Якщо якісь комірки виділені, заміна відбудеться лише у них, щоб не зачепити інші значення.

A	B	C	D	E	F	G	Н	1	J
Назва і ЄДРПОУ підприємства	Прибуток								
КП "Щось і Дещо";656565	10000								
ТОВ "Підприємство";232323	1 000 000				Найтии	заменить		?	×
КП "Супер-комунальники";555555	65 тис.					Jamentino			
		Найти	замен	ить					
_		Найти	· F						Y
-		Заменя	ить на:	;		Microso	oft Excel		
		Замени	пъ <u>в</u> се	Заменит	•	Тоиск заверш	іен; выполне К	но 3 замен.	

З'явились крапки з комою. Тепер розділимо колонку на дві:

4	айл Главная Вставка Разметка с	траницы	Формулы Да	нные		Рецензирова	ние	Вид		$\frown$			
	Из Ассезя Из Интернета Из Интернета Из текста Получение внешних данных А1 ▼	Обнова все т зва і ЄДР	<ul> <li>Подключени</li> <li>Свойства</li> <li>Изменить свя</li> <li>Подключения</li> <li>ПОДключения</li> </ul>	а 3И а	\$↓ ⊼↓	АЛ Сортировка Сорт	Фильтр Фильтр ировка и	🖗 Очистить 🍒 Повторита У Дополнит Фильтр	ельно	екст по Уда олбцам дибл	- 📑 икаты д Ра	роверк анных бота с д	а Консолид занными
	A	В	С		D	F	F	6	н		1		К
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18	нава I «ДРПСУ підприкита» М' ПЦСо I Десс"55555 ТОВ "Підприкита";523323 КП "Супер-комунальники";55555		Прибуток 10000 1000 000 65 тис.	Да Ecr У	нные ти это орма <sup>1</sup> (кажи (кажи (кажи (кажи (кажи (кажи)) (кажи) (кажи)) (кажи) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (кажи)) (	восприняты к з верно, наже и коодных да те формат да с разделител фиксировани аврительный п ава 1 сдряс "Пося 1, а в "Підприев в "Підприев	Масте ак список пе кнопк: нных: ян: яных: ян: ян: ян: ян: ян: ян: ян: ян: ян: ян	р текстов () значений с рак "Далее >", в и - значения и - значения и - поля ичес вобранных дан искоства (565 32323 вох"; 555555	разбор) зделителян противном полей отде от заданнум ных:	- шаг 1 из як. случае укажи аляются знака ко ширину	з те форм ем-разде	?	605. PH
19 20								Отнена	< Has	хад Да	ūee >	D	08010

А) Спочатку звільніть місце для колонок, які з'являться в результаті поділу, інакше вони зітруть значення в наступних стовпцях. Якщо не знаєте, скільки колонок отримаєте в результаті, краще перенести стовпчик з багатьма змінними в кінець таблиці.

Б) У меню «Дані» виберіть «Текст по стовпчиках». Зараз ми маємо текст з розділювачами, тож обираємо цю опцію в меню текстів по стовпчиках. Якщо ж кожне значення займає однакову кількість знаків, можна використовувати формат даних «фіксованої ширини».

А	В	С	D	E	F	G	Н	1	J		K
Назва і ЄДРПОУ підприємства		Прибуток			Мастер	текстов (	разбор) -	шаг 2 из	3	?	
КП "Щось і Дещо";656565		10000	Данный д	циалог позво	оляет установ	зить раздел	ители для те	екстовых дан	ных. Резул	ьтат	
ТОВ "Підприємство";22223 КП "Супер-комунальники";555555		1 000 000 65 тис.	Выводити Символо У знак Точн запя проб	ся в окне об он-разделит к <u>т</u> абуляции ка с запятой ттая бел гой:	разца разборі елен являетс ]	a. R: Qrpa+	цтать после, инчитель стр	довательные ок: "	е разделите	ли одн	ин
			Образец Назва КП "Пр ГОВ "1 КП "С;	і ЄДРПОУ ось і Дещ Підприємс упер-кому	підприємо о" тво" нальники"	5565 2323 5555	65 23 55				<b>^</b>
						Отмена	< <u>H</u> asa	да Да	nee >	<u>[</u> 071	080

В) Натиснувши «Далі», виберемо розділювач. У нас крапка з комою. У попередньому перегляді бачимо, що дані розділились на дві колонки. Якщо вибраного розділювача немає серед вказаних, можете самі ввести його в полі «інший». Майже готово, натискаємо «Далі».

1	А	В	С	D	E	F	G	н	1	J	K
1	Назва і ЄДРПОУ підприємства		Прибуток			Мастер	текстов (	разбор) -	- шаг 3 из	3	? ×
2	КП "Щось і Дещо";656565		10000								_
3	ТОВ "Підприємство";232323		1 000 000	данный д	иалог позв	оляет устан	овить для ка	кдого столо	ца формат да	HHBIX.	
4	КП "Супер-комунальники";555555		65 тис.	Формат,	данных сто	поца	"O6		гса наиболее	VHUREOCARL	HUM (100
5					ии		значений эт	ого формат	а осуществля	ется автом	атическое
6							преобразов всех прочих	ание числов « значений -	ых значений в текст.	в числа, да	т - в даты, а
7					a: <u>H</u> en	×			Подробнее		
ð o					пустить сто.	юец					
9 10				Поместит	ь в: \$А\$1						1
11				Образец	разбора да	нных					
12						_					
13				Общий			06mrd	ž			
14				Назва КП "Ще	і ЄДРПОУ всь і Деш	підприєї о"	65656	65			
15				TOB "I	Підприеме		23232	23			
16				ku "oj	/пер-кому	нальники.	55555	55			~
17				<							>
18											
19							Отмена	< <u>H</u> asa	ад Дал	iee >	<u>Г</u> отово
20						-			-		

Г) За бажанням можемо вибрати формат колонок, який хочемо отримати в результаті. Встановимо «Текст» для колонок назви та ЄДРПОУ. Якщо залишимо колонку з ЄДРПОУ як число, втратимо нулі на початку коду. Взагалі, якщо дані, записані числами, є кодом, ідентифікатором об'єкта (як-от код ЄДРПОУ), а не призначені для здійснення математичних розрахунків (наприклад, прибуток) — обирайте текстовий формат для колонки. Натисніть «Готово».

	A	В	С
1	Назва підприємства	єдрпоу	Прибуток
2	КП "Щось і Дещо"	656565	10000
3	ТОВ "Підприємство"	232323	1 000 000
4	КП "Супер-комунальники"	555555	65 тис.
5			

Залишилось встановити назву, і маємо дві окремі колонки, по одній змінній в кожній.

Змінна «Прибуток» записана не уніфіковано. Необхідно перевести дані в числовий формат. Для того, щоб було зручно працювати, оформимо дані як таблицю.

X	Файл Главн	-   <del>-</del> ая Встав	ка Размет	ка страниць	Формул	я Дан	ные і	Рецензиров	Кн ание Ви	ига1 - N u	licrosoft Ex	cel			
	Вставить →	Calibri Ж.К.Ц	• 11 •   ⊞ •   3	• A* A*	= = =		िंग Пер	енос текста ъединить и	поместить в	центре *	Общий 🥶 - % (	•	Усло	вное	Форматироват
Бу	/фер обмена 🕞		Шрифт	5			Выравни	вание		Gi.	Числ	o 5		(	ТИЛИ
-	C4	• (*	$f_{N}$	65 тис.											
	da .	А		В	C		D	E	F	G	Н	1	J	К	L
1	1 Назва підпр	иємства		єдрпо	У Прибуто	ж									
2	2 КП "Щось і Д	Зещо"		6565	565	10000									
3	3 ТОВ "Підпри	емство"		2323	323 1 (	000 000									
4	4 КП "Cynep-н	омунальни	ики"	5555	55 65 тис.										
	5														

В головній вкладці виберіть «Форматувати як таблицю» та оберіть будь-який дизайн.

	A	В	С
1	Назва підприємства 🛛 🔻	ЄДРПОУ 🔻	Прибуток 🔽
2	КП "Щось і Дещо"	656565	10000
3	ТОВ "Підприємство"	232323	1 000 000
4	КП "Супер-комунальники"	555555	65 тис.
5			

В такій таблиці зручніше відбирати дані за значенням колонок та вираховувати нові змінні.

Тепер зробимо запис суми прибутку однаковим. Спершу виділимо колонку і оберемо формат комірок «Числовий»:

🗶 🛃 🗉 - (ല - 💷	Книга	1 - Microsoft E	xcel			PaGo	та с таблицами	
Файл Главная Вставка Разметка	страницы (	Формулы Дані	ные Рег	цензирован	ие Вид	К	онструктор	
Calibri • 11 •	A ^ =	= 🕳 🗞 -	📑 Перен	ос текста			Числовой	Ţ
Вставить 🛷 ЖКЦ - 🔛 - 🆄	• <u>A</u> • 🔳	= =   ∉ ∉	-ат Объед	инить и по	местить в це	нтре т	<b>∰</b> ~ % 000	\$,0 ,00 ,00 \$,0
Буфер обмена 🖙 Шрифт	G.		Выравнива	ние		- G	Число	
C1 • fx Π	рибуток						$\sim$	
A	В	С	D	E	F	G	Н	1
1 Назва підприємства	ЄДРПОУ 👻	Прибуток 🔽						
2 КП "Щось і Дещо"	656565	10000,00						
3 ТОВ "Підприємство"	232323	100000,00						
4 КП "Супер-комунальники"	555555	65 тис.						

Бачимо, що значення з пробілами стали числами, а от «65 тис.» так і залишились текстовим рядком. В таблиці з трьома рядками можна це виправити вручну. Але уявіть, що таких рядків - кілька тисяч. Відфільтруємо значення, що містять «тис»:



Виділимо колонку з цими значеннями і замінимо «тис.» на значення порожнього рядка (дві послідовні лапки, ""), або залишіть поле «Замінити на» повністю порожнім. Отримали 65. X 🚽 🗉 - (2 - 1 -Книга1 - Microsoft Excel Файл Главная Буфер обмена 😘 Calibri - 11 - A A 3 📆 - % 000 🔤 *f*<sub>x</sub> Прибуток C1 ж К 🗏 🌭 - 🔬 - 🖂 - % 🖑 🛷 👗 Вырезать • Єдрпоу • прибут Назва підприємства КП "Супер 😤 Пара Ê Специальная вставка... <u>В</u>ставить 9 <u>У</u>далить 10 11 12 Очистить содера Формат дчеек... Ширина столбца.. 14 ⊆крыть Показать 15

Тепер скопіюємо ці значення в нову тимчасову колонку:

Залишилось помножити значення на тисячу. Поставте курсор у першу комірку колонки з відфільтрованими значеннями, і введіть формулу:

=[(	@Столбец1]	*1000		
	В	С	D	
-	ЄДРПОУ 👻	Столбец 👻	Прибуток 🗐 🗐	
	555555	65,00	бец1]*1000	Į
	=[( 	=[@Столбец1] В ▼ €ДРПОУ ▼ 5555555	=[@Столбец1]*1000 В С • ЄДРПОУ • Столбец • 555555 65,00	=[@Столбец1]*1000 В С D • ЄДРПОУ • Столбец • Прибуток • Г 555555 65,00 бец1]*1000

Поставте знак «=» і почніть вводити свою формулу. Назви колонок писати непотрібно, достатньо просто клікнути на потрібний стовпчик.

Приберемо фільтр — і маємо чисту таблицю, щоправда, з зайвим стовпцем.

	А	В	С	D
1	Назва підприємства 📃 🔻	ЄДРПОУ 👻	Столбец 🔻	Прибуток 💌
2	КП "Щось і Дещо"	656565		10000,00
3	ТОВ "Підприємство"	232323		1000000,00
4	КП "Супер-комунальники"	555555	65,00	65000,00
5				

11

Якщо ми відразу видалимо цей стовпчик, обраховане значення зникне, адже воно існує як результат арифметичної операції. Щоб уникнути цього, виділіть колонку, яку слід зал ишити, скопіюйте її і вставте на місце тимчасового стовпця.

🗶   🛃 🖤 • (° •   =			Книга	1 - Microso	oft Excel	
۰	айл Главная	Вставка Разметка	страницы	Формулы	Данные Реце	ізирование В
Вс	Гарить 🛷 Ка	libri • 11 • K K Y • ⊞ •   🌺	A <sup>*</sup> A <sup>*</sup> ≡	= <mark>=</mark> »·	🛱 Перенос 🗱 Объедии	текста нить и поместить
Буф	ер обмена 🗔	Шрифт	6		Выравнивани	e
	C1		рибуток			
1		A	В	С	D	E I
1	Назва підприє	Morea	єдрпоу -	Прибуто 🗠	Прибуток2 🗾	
2	КП "Щось і Деи	цо"	656565	10000,00	10000,00	
3	ТОВ "Підприєм	ство"	232323	1000000,00	100000,00	
4	КП "Супер-ком	унальники"	555555	65000,00	65000,00	
5					Ctrl) -	
6					Вставка	
7					🗎 fx %jx	<b>*</b>
9					m 🙃 📾	_
10						
11					Вставить значен	RN
12					123 123 123	
13					Другие параме	гры вставки
14					%	
15						

В налаштуваннях вставки слід обрати «Значення» або «Значення і формати». Тоді значення комірок перестануть залежати від формули, за допомогою якої вони були отримані. Залишилось видалити непотрібну колонку, і таблиця чиста.

	А	В	С
1	Назва підприємства 📃 🔻	ЄДРПОУ 👻	Прибуто 💌
2	КП "Щось і Дещо"	656565	10000,00
3	ТОВ "Підприємство"	232323	100000,00
4	КП "Супер-комунальники"	555555	65000,00
5			

В Excel (без застосування макросів — програмування) можна обраховувати значення, видаляти рядки, що повторюються, відкривати формати відкритих даних .csv, .xml, підсумовувати дані в звідних таблицях і звідних діаграмах, відбирати дані за умовами. Проте функції Excel обмежені, і більше Ви можете зробити за допомогою інструментів, про які написано в секції «Чим почистити таблицю, якщо Excel'я недостатньо». Інструменти, про які йдеться, потужні та безкоштовні.

# Формати

# Табличні

Дані у формі рядків та колонок. Ці формати можна відкрити як у середовищах програмування, так і у звичних табличних редакторах — Excel, Libre Office Calc тощо.

#### .xls / .xlsx

Формат даних Excel, який можна прочитати у більшості табличних редакторів

- + Легко створити
- + Можливо прочитати в середовищах програмування

Перестає бути машиночитаними відкритими даними, якщо порушена структура рядків і стовпців (заголовки замість рядків всередині таблиці, об'єднані комірки тощо).

Інформація, закодована у форматуванні (колір комірки, шрифт і т. ін.) незрозуміла машині. Тому уникайте цього в остаточних документах.

#### .csv, .tsv

Простий текст з розділювачами, які позначають структуру таблиці. Відповідно, є розділювач стовпців та рядків таблиці. Найпоширеніший формат .csv, де стандартно

- колонки розділені комами
- рядки розділені знаком нового рядка \n
- значення колонок в подвійних лапках (аби коми всередині тексту не руйнували структуру файлу)
- нецілі числа розділені крапками: 12.345
- кодування UTF-8

Якщо Ви вносите зміни в цей формат, необхідно попереджати про це користувача даних!

Також поширений формат .tsv, в якому колонки розділені знаком табуляції \t. Зручний для текстових даних з комами.

При збереженні файлу у форматі CSV, Excel за замовчуванням зберігає файл .csv у кодуванні вашої операційної системи. Це, зокрема, впливає на десятковий розділювач та розділювач стовпчиків. Для локалізованих українською та російською мовами версій MS Office це кодування Windows-1251, в якості розділювача виступає крапка з комою «;», а десят-

кового розділювача — кома. Часто Excel спотворює дати або невірно читає числа з десятковими значеннями. Краще скористатись Libre Office Calc або Google Таблицями, аби зберегти файл в загальноприйнятому форматі .csv: UTF-8, коми розділяють стовпці, текстові значення в подвійних лапках, десятковий розділювач — крапка. Конвертувати таблицю з Excel y .csv можна і за допомгою онлайн-конвертерів (наприклад, https://convertio.co/ru/xlscsv/).

# Ієрархічні

Ці формати дозволяють вийти за межі структури колонок і рядків. Наприклад:

Колонка 1	Колонка 2	Колонка 3
A	Б	Г
A	Б	Д
A	В	E
A	В	E

В таблиці ієрархічні дані призводять до багаторазового повторення значень, а при обробці даних часто треба застосовувати складні фільтри.

Представимо цю ж таблицю в форматі JSON – JavaScript object notation:

JSON можна прочитати в більшості програмних середовищ. Цей формат чудово підходить для передачі даних через Інтернет, зокрема, за допомогою API. Якщо у вас великий масив ієрархічних даних — обирайте JSON. Зверніть увагу, як суттєво зменшився обсяг тексту в цьому форматі, порівняно з таблицею. Завдяки ієрархічній структурі такі формати дозволяють зменшити розмір файлу з ієрархічними даними. Незважаючи на те, що JSON відносно легко читати і редагувати людині, цей формат даних ліпше створювати машинним шляхом. Наприклад, багато баз даних дозволяють експортувати дані у JSON, втім, як і в інші відкриті формати. Якщо Ви редагували JSON вручну, перед публікацією варто перевірити файл на наявність помилок. Для цього є багато безкоштовних сервісів<sup>2</sup>, деякі з них дозволяють і зручно редагувати файл. Дані в JSON мають бути у кодуванні UTF-8, як того вимагає стандарт RFC 7159<sup>3</sup>), аби не спотворити символи, і щоб файл легко було прочитати на інших комп'ютерах.

На малюнку показано, як відображається JSON в кодуванні Windows-1251. Знаки заміщення тут з'явилися замість українських літер — не робіть так. Ви-користовуйте UTF-8.

1	<b>॑{!���!: 'json'</b> ,
2	<b>॑'�� ���'</b> : [
3	{ <b>'*********</b> 1': 1},
4	{ <b>***********</b> 2; 2}
5	

#### .xml

Ієрархічний формат даних, зручний для передачі через Інтернет. Дані організовані в елементи, які можуть містити інші елементи — і так далі. Приклад попередньої таблиці:



<sup>2</sup> http://jsonviewer.stack.hu/, http://jsoneditoronline.org/

<sup>3</sup> https://tools.ietf.org/html/rfc7159

Формат XML схожий на структуру тегів HTML на веб-сторінці. Так само, при роботі з XML слід перевіряти кодування (UTF-8) і впевнитись у відсутності помилок перед публікацією даних.

Якщо Ви маєте просту прямокутну таблицю, немає сенсу зберігати її в ієрархічних форматах. Тож якщо Ви не певні, який же формат використати — обирайте .csv.

# Як зберегти таблицю в форматі відкритих даних?

Ми підготували порівняльну таблицю найпопулярніших програм для роботи з даними і показали їх можливості експорту в інші формати:

	csv / tsv	JSON	XML	Примітки
Excel				.csv лише з розділювачем «;» в локальному кодуванні. xml теж лише у локальному кодуванні, використати всі можливості цього формату чарез Excel не вдасться.
LibreOffice Calc				Опція «редагувати налаштування фільтра» у вікні «Зберегти як» для вибору кодування і розділювача в .csv
Google sheets				Опція «Завантажити як» + можли- вий онлайн-доступ до даних у сsv: функція «Файл - Опублікувати як»
OpenRefine				Опція «Export»

У розглянуті формати можна зберегти бази даних і витяги з них, а також дані з програмних середовищ. Як бачите, майже всі програми працюють з табличними форматами, хоча Excel зберігає навіть табличні дані у нестандартному форматі (перед публікацією можна перетворити на класичний csv в LibreOffice Calc). Для створення ієрархічних форматів краще послуговуватись більш просунутими інструментами.

# Геопросторові дані

Це особливі формати даних, які використовуються для картографії та геопросторового аналізу. Такі дані обов'язково містять географічні координати об'єктів: точок, ліній та полігонів, а також їхні атрибути — дані, прив'язані до географічних форм. Точка — це пара координат широти і довготи, що позначає конкретне місце на площині. Лінія — це дві точки, отже, відрізок. Полігони — це множини точок. Полігони можуть бути замкнені: в такому випадку це фігура — наприклад, межі Вашого міста на карті. У незамкненому полігоні точка кінця і початку не має збігатися — наприклад, як у координатах вулиці, яка починається в одному місці, а закінчується в іншому.

Отримання географічних координат об'єктів називається геокодуванням. У Вас є список адрес і Вам потрібно отримати їхні координати? Шукати координати кожної точки на сервісі карт вручну — не найкраще рішення. Без програмування геокодувати адреси можна у QGis, також є додаток для Google Таблиць4. Щоб легко і без помилок геокодувати адреси, вони мають бути однакового формату. Якщо програма геокодування не знаходить координати місця, зміна формату адреси може допомогти.

## Шейпфайли

Це набір з файлів .shp, .shx, .dbf, які містять координати форм та дані атрибутів. По суті, шейпфайл — це не один файл, а декілька. Наприклад, форма — це район у області або населеному пункті, отже, координати полігона. А атрибути — дані про цю форму: назва, скажімо, чисельність населення і все, що Ви бажаєте відобразити в даних. Або форма — точки-координати шкіл і адреси у вигляді атрибутів.

QGis<sup>5</sup> — програма з відкритим кодом (безкоштовна і підтримується спільнотою користувачів), у якій можна створювати, читати, редагувати шейпфайли й інші геопросторові формати. Геопросторові дані місць з усього світу збирає і публікує проект Open Street Map. Там можна знайти і завантажити координати адміністративних одиниць, об'єктів інфраструктури, житлових будинків тощо.

Комерційний стандарт для роботи з геопросторовими даними — програма ArcGis. Вона має більше функцій, проте є платною. Також з географічними форматами можна працювати в середовищах програмування. Відповідні додатки є для R, Python та інших мов.

# GeoJSON

Це розширення формату JSON для передачі і зберігання географічних даних. Кожному елементу в GeoJSON присвоюються значення координат форми («geometry»), що містить тип форми (точка / лінія, як з двома, так і більше парами координат / полігон, замкнений / а також множини цих форм) та дані-атрибути кожного об'єкта («properties»). Шейпфайли легко перетворити в GeoJSON, і навпаки. Це можна зробити в QGis або іншій подібній програмі.

<sup>4</sup> https://chrome.google.com/webstore/detail/geocode-by-awesome-table/cnhboknahecjdnlkjnlodacdjelippfg

<sup>5</sup> http://www.qgis.org/uk/site/forusers/download.html

Як і JSON, GeoJSON чудово підходить для передачі даних мережею Інтернет і може використовуватись для API. В роботі з GeoJSON варто дотримуватись тих самих правил, що і при роботі з JSON. А саме: перевіряти файл на наявність помилок перед публікацією і дотримуватись формату UTF-8.

#### Чим почистити таблицю, якщо Excelя недостатньо

Змінити структуру таблиці можна багатьма способами. Щоб не переставляти комірки таблиці руками, можна використати Open Refine<sup>6</sup> або пакет tidyr<sup>7</sup> для мови програмування R (базових навичок програмування вистачить), Pandas для мови Python, макроси тощо. Без програмування з чисткою даних допоможе впоратись OpenRefine, безкоштовна програма з відкритим кодом.

Якщо в своїй роботі Ви постійно виконуєте одну й ту саму дію багато разів вручну, швидше за все, це легко можна зробити за допомогою програмування. В результаті Ви заощадите свої зусилля і навчитесь чомусь новому та цікавому. Для кожного з вказаних інструментів є докладна документація, онлайнпосібники, безкоштовні курси і спільнота користувачів, які допомагають одне одному.

# **Open Refine**

Унікальна функція OpenRefine — це **виявляти різні варіанти написання однієї** назви (кластеризувати текстові дані) чи дати та зводити їх до одного вигляду. Дуже корисно для виявлення друкарських помилок при ручному введенні даних.

Окрім цього, OpenRefine може виконувавати функції Excel'я чи LibreOffice на набагато більших даних, зокрема:

- повертати дані із рядків у стовпці і навпаки (транспонувати таблиці);
- знаходити і виправляти помилки у наборі даних;
- розділяти дані з однієї колонки на декілька;
- здійснювати прості арифметичні операції;
- об'єднувати кілька наборів даних в один;
- конвертувати в формати відкритих даних (функція «Export»).

<sup>6</sup> http://openrefine.org/

<sup>7</sup> http://tidyr.tidyverse.org/

Посібник підготовлено у рамках проекту USAID/UK Aid «Прозорість та підзвітність у державному управлінні та послугах», який реалізує Фонд «Євразія», у партнерстві з ГО «Агенція журналістики даних».









Державне агентство з питань електронного урядування України

TEXTY.ORG.UA