Відкриті дані для міст

Практичний аспект

КИЇВ • 2019

Автори:
Андрій Савчук
Руслана Величко

Дизайн, верстка — Іван Юрчик

*Посібник **"Відкриті дані для міст. Практичний аспект"** – це настільна книга для органів місцевої влади та активістів, для тих, хто вболіває за відкриті дані для своєї громади. Він містить практичні рекомендації щодо етапів впровадження політики відкритих даних. Посібник є дороговказом та дасть відповіді на запитання, як саме організувати роботу з оприлюднення даних у машиночитних форматах, яка структура даних є прийнятною тощо*

Відкриваючи ці дані для загального доступу, ми створюємо новий світ можливостей для влади, бізнесу, громади, установ та організацій

Підготовку та друк здійснено за підтримки National Endowment for Democracy
у рамках ініціативи "Дані міст"

ВСТУП

Шановні Читачі,

з радістю передаємо у ваші руки цей посібник, створений у рамках ініціативи "Дані міст" з метою популяризації руху відкритих даних!

Дані – це ресурс, яким потрібно ділитися. Адже разом ми можемо більше, ніж наодинці. Публікація "Відкриті дані для міст" – це не тільки корисне зібрання порад, досвіду та рішень, але й розгорнута інструкція на тему впровадження політики відкритих даних.

Як? Де? Чому? Видання містить корисні та дієві рекомендації у сфері чистки, вибору формату, публікації та аудиту наборів даних. Книга розрахована на широкий загал, але стане в нагоді та буде особливо корисною для органів влади, журналістів та громадських активістів, які працюють з даними.

Відкриті дані потрібно використовувати не лише задля виконання покладених державою обов'язків, але й для прийняття рішень. Дані можуть допомагати та показувати, де у місті найболючіша зона, що потребує якнайшвидшої допомоги (дані щодо матеріально-технічної бази навчальних закладів, медичних закладів, дані щодо швидкості реагування швидкої медичної допомоги по районах, дані щодо району, де населення найбільше користується соціальною допомогою тощо). І якщо зрозуміти та усвідомити, якою силою та яким важливим інструментом є відкриті дані, то можна в рази поліпшити якість управління містом, дізнатися більше про населення та його потреби, передбачити та допомогти вирішити ситуації, перш ніж вони стануть проблемою.

Потенціал відкритих даних – це в першу чергу інвестиція у майбутнє, адже в подальшому дані можуть принести не тільки відкритість та прозорість для міста, але й зменшити кількість запитуваної інформації, що значно скоротить час для обробки таких запитів. Наприклад, у місті Гданськ (Польща) всі запити на публічну інформацію та відповіді на них публікуються в режимі реального часу. Дані можуть використовуватися не тільки для контролю, але й для створення зручних умов жителям міста. Якщо публікувати дані у правильному машиночитному форматі, то на основі цих даних можна створити різноманітні сервіси, починаючи з інформації про школи, лікарні та завершуючи найбільш безпечними маршрутами від роботи до домівки, включаючи інформацію про якість повітря. Дані можуть передавати інформацію та продемонструвати, яка сфера зайнятості найбільш оплачувана у вашому місті, а яка найбільш запитувана, яка послуга для жителів на сьогодні є найбільш затребуваною та необхідною тощо. Як висновок – дані можуть допомогти знайти відповідь на багато існуючих та гострих запитань.

Але для подібного ефекту органи місцевого самоврядування повинні публікувати дані так, щоб їх можна було легко використовувати та мати змогу для автоматизованої обробки.

Ми сподіваємося, що у Вас вже є розуміння щодо важливості відкритих даних міст. У цьому посібнику ми намагатимемося поглибити знання більш практичними та наочними порадами, які допоможуть зробити Ваше місто прозорішим та комфортнішим.

Бажаємо приємного користування!

Команда “Дані міст”

ДАНІ

Що таке дані

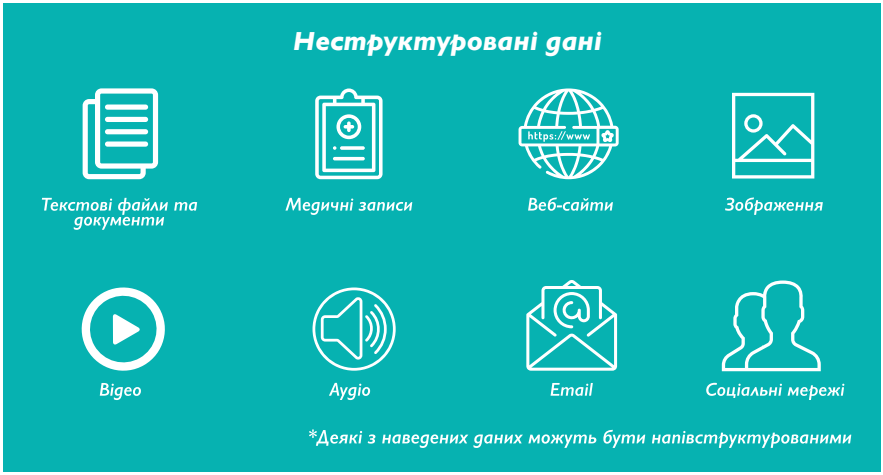
Дані *(англ. Data) – один із термінів, точне значення якого досить проблематично визначити. Тому у простому розумінні дані – це інформація, якою можна поділитись і яку можна обробити, в першу чергу машинним (автоматичним) способом. Крім того, така інформація повинна нести у собі певний зміст (значення).*

Для цілей даного посібника, цілком можна стверджувати, що дані – це інформація, яка певним чином характеризує якийсь об'єкт, передає його властивості, відомості чи показники. Іншими словами – допомагає зрозуміти те, що описує.

Зазвичай дані об'єднують у набори даних (датасети), які найчастіше записують у вигляді таблиць чи баз даних. З точки зору українського законодавства, набір даних — це сукупність однорідних значень (записів) даних та метаданих, що їх описують.

Загалом дані можна поділити на структуровані та неструктуровані. Деякі дослідники також окремо виділяють напівструктуровані дані, які є комбінацією обох попередніх типів (наприклад, XML).

- **Структуровані** – це добре впорядковані дані, які організовані та описані за певною моделлю (стандартом). Більшість програм працюють саме із цим типом даних.
- **Неструктуровані** – це інформація, яка не має попередньо визначеної моделі даних. Така інформація зазвичай є текстовою. За підрахунками експертів, неструктуровані дані складають близько 80% від усіх світових даних.



Зазвичай неструктуровані дані можна перетворити у структуровані. Візьмемо, до прикладу, інформаційну довідку про результати голосування у міській раді міста "А".

> *Сьогодні, 17 лютого 2018 року, депутати нашого міста прийняли кілька важливих рішень. Зокрема, за проект ухвали про перейменування вулиці Леніна на вулицю Тараса Шевченка проголосувало 37 депутатів, за проект рішення №123 "Про перелік відкритих даних" – 45. Також міська рада 38 голосами "за" ухвалила рішення щодо створення скверу на перетині вулиць Хмельницького та Івана Франка.*

Таку ж інформацію можна подати у вигляді структурованих даних, тобто таблиці:

date	id	description	votes
2018-02-17	122	Щодо перейменування вулиці Леніна	37
2018-02-17	123	Про перелік наборів відкритих даних	45
2018-02-17	124	Щодо створення скверу у Франківському районі	38

У другому випадку, дані можна легко об'єднати із такою ж інформацією з інших сесій міської ради і використати для аналізу чи візуалізації.

***ВАЖЛИВО!** Не всі дані, що містяться у таблицях є структурованими! Зокрема, не є такими скан-копії документів та таблиці із неправильною структурою (про ці аспекти поговоримо у наступних розділах).*

БАЛАНС

на 01 січня 2019 року

Форма №1-дс

АКТИВ	Код рядка	На початок звітного періоду	На кінець звітного періоду
1	2	3	4
I. НЕФІНАНСОВІ АКТИВИ			
Основні засоби:	*1000*	*53314477*	*52977102*
первісна вартість	*1001*	*55468732*	*55506450*
знос	*1002*	*2154255*	*2529348*
Інвестиційна нерухомість:	*1010*	-	-
первісна вартість	*1011*	-	-
знос	*1012*	-	-
Нематеріальні активи:	*1020*	-	-
первісна вартість	*1021*	-	-
накопичена амортизація	*1022*	-	-
Незавершені капітальні інвестиції	1030	-	-
Довгострокові біологічні активи:	*1040*	-	-
первісна вартість	*1041*	-	-
накопичена амортизація	*1042*	-	-
Запаси	1050	110003	85879
Виробництво	1060	-	-
Поточні біологічні активи	1090	-	-
Усього за розділом I	*1095*	*53424480*	*53062981*

Які бувають дані

Крім поділу на структуровані та неструктуровані, у світі існує і ціла низка класифікацій даних. Так, відповідно до інформації, які несуть певні файли, зазвичай виділяють:

- текстові;
- табличні;
- графічні;
- аудіо;
- відео;
- геопросторові;
- архівні та інші дані.

З точки зору, статистики, можна визначити наступні рівні даних:

- глобальний,
- державний,
- місцевий.

У нашому посібнику ми здебільшого розглядатимемо саме останній рівень, який проте є одним з найважливіших. Адже майбутнє – це світ міст, комфортних та затишних; міст, які безперервно розвиваються, де влада, бізнес та громадяни постійно взаємодіють між собою; міст, де рішення приймаються стратегічно, виважено та цілеспрямовано. Бо саме такого міста прагнуть його мешканці та справжні управлінці.

А досягти ефективності роботи міста можна тільки завдяки роботі з даними. Планування, прогнозування, прийняття рішень – усе має відбуватись на основі детального аналізу фактів та даних. Але дуже часто трапляється так, що міська рада не в змозі самостійно якісно обробити ту чи іншу інформацію. Саме тому, нею потрібно ділитися з усіма зацікавленими сторонами. Адже як показує практика, інформація, яка є в розпорядженні міської ради дуже цікава для мешканців.

Так, Фонд TechSoup у співпраці з Громадянською мережею ОПОРА в рамках ініціативи "Дані міст" проводив опитування щодо даних, які варто відкрити в містах. Відповідно до результатів, мешканці українських міст найбільше потребують інформації про:

- екологічний стан (якість повітря, води);
- фінансово-господарські дані (бюджет, витрати, тендери тощо);
- генеральний план міст, картографічна інформація про місто;
- комунальне майно (вільне, оренда тощо);

- комунальні підприємства і послуги, які ті надають (тарифи, інвестиційні програми та звіти щодо них);
- освіта;
- охорона здоров'я;
- вивіз сміття, утилізація;
- криміногенна ситуація в місті;
- громадський транспорт.

Звісно, це неповний перелік даних, адже кожне місто унікальне. Тому інформація, яку варто відкрити насамперед, може різнитися. Отже, якщо Ви хочете відкрити дані та бажаєте ефективно використати й так обмежені ресурси, треба спитати місцевих мешканців:

- які проблеми їх турбують;
- які вони бачать технічні рішення, що могли б вирішити ці проблеми;
- які дані потрібно для цього відкрити.

Можливо, дані взагалі не потрібно відкривати? Чи потрібно? Давайте про це поговоримо у наступному розділі.

ВІДКРИТІ ДАНІ

Що таке відкриті дані

Відкриті дані *(англ. Open Data) – це система поглядів, яка відображає ідею про те, що певні дані мають бути вільно доступними для обробки машинним способом та подальшого використання і розповсюдження без жодних обмежень і контролю, в тому числі з комерційною метою. Для звільнення даних від обмежень авторського права можна використовувати так звані вільні ліцензії, зокрема Creative Commons license.*

Відкриті дані – це в першу чергу можливість легкого, ефективного та стандартизованого обміну даними між усіма зацікавленими сторонами, в тому числі самими органами влади. Це дієвий механізм, який не лише "змушує" ділитися інформацією, але й допомагає отримувати дивіденди від цього.

Історія відкритих даних

Ідея[1] загального суспільного блага від відкритості даних не є новою. Проте спершу вона стосувалась результатів наукових досліджень і була висловлена ще у 1942 році. Її автором став американський соціолог, один із засновників школи структурно-функціонального аналізу **Роберт Кінг Мертон**. Теорія, що носить його ім'я, демонструє важливість того, що результати дослідження повинні бути доступними для всіх. Мертон вважав, що кожен дослідник повинен здійснити свій внесок до "загального банку" і відмовитися від прав інтелектуальної власності в ім'я прогресу знань.

[1] *http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data*

Сам термін "відкриті дані" вперше з'явився у 1995 році у документі[2] Американського наукового товариства і стосувався розкриття геофізичних та геологічних даних, зокрема їхнього "повного та відкритого обміну".

Вже сучасні принципи, які сьогодні дозволяють нам визначати і оцінювати відкриті дані, були сформульовані у 2007 році на зустрічі теоретиків та активістів Інтернету, що відбулась поблизу Сан-Франциско у Себастополі[3].

Одними з авторів цих принципів стали **Тім О'Рейлі** – активіст руху за вільне програмне забезпечення з відкритим вихідним кодом, один з головних ідеологів Web 2.0 та **Лоуренс Лессіг** – професор права Стенфордського університету та засновник організації Creative Commons.

Ідея, що лежить в основі даних принципів, полягає в тому, що публічні дані – це спільна власність, якою можна ділитися і яку можна використовувати. По суті, це ті ж самі принципи, що лежать в ідеології програмного забезпечення з відкритим вихідним кодом: *відкритість, участь та спільна робота.*

Можливо, у 2007 році пропоновані принципи звучали як утопія, але вже за рік президент США Барак Обама підписав три президентських меморандуми, що стосувалися прозорості даних і відкритого уряду. Відтоді озвучені ідеї почали ставати реальністю. У 2009 році був запущений урядовий портал відкритих даних США[4] – *data.gov*, за рік ініціатива стала дійсністю у Великій Британії[5] – *data.gov.uk*. В Україні аналогічний веб-ресурс *data.gov.ua* був запущений[6] у 2015 році.

Використання відкритих даних

Відкриті дані допомагають владі приймати ефективніші рішення, оскільки до цього процесу долучаються мешканці міста. Інфраструктура відкритих даних створює сприятливі умови для ефективного громадського контролю, що покращує підзвітність міської влади громаді, підвищує прозорість. Це, в свою чергу, позитивно впливає на ефективність управлінських рішень, а отже – на якість політики, інвестиційну привабливість тощо.

Віддавши дані якісно, місто створює для громадян умови, коли ті можуть починати самостійно вирішувати ідентифіковані проблеми за допомогою цих даних, а для бізнесу — покращувати якість послуг та товарів, знаходити клієнтів, партнерів.

[2] *https://www.nap.edu/read/18769/chapter/1*

[3] *https://opengovdata.org*

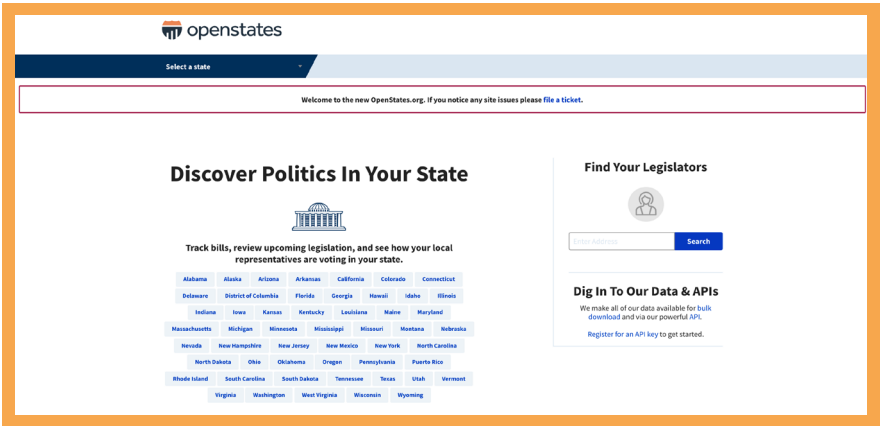
[4] *https://www.data.gov/about*

[5] *https://data.gov.uk/about*

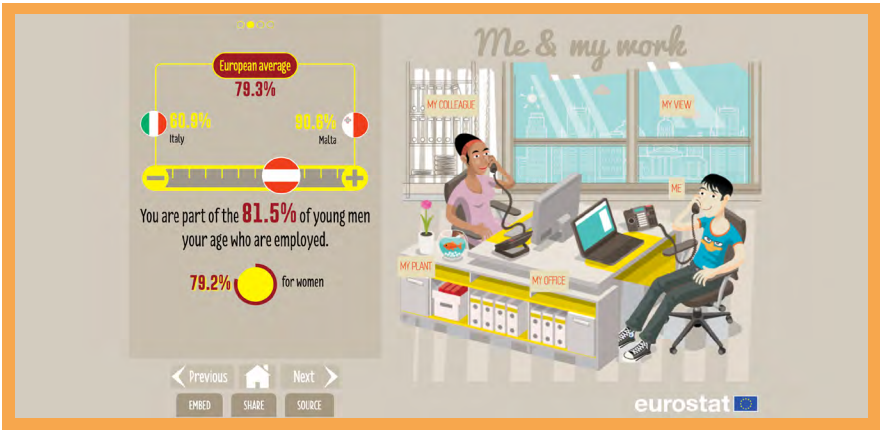
[6] *https://data.gov.ua/pages/about*

Саме тому сьогодні у світі існують сотні сервісів на основі відкритих даних. Давайте розглянемо деякі з них.

OpenStates[7] – сервіс, що об'єднує законодавчу інформацію з 50 штатів США, округу Коламбія та Пуерто-Ріко. Портал дає змогу ознайомитися зі змістом кожної з ініціатив. Характерною особливістю ресурсу є можливість порівняння регулювання окремих галузей в різних штатах. Також сайт дає змогу знайти депутата, який представляє Ваш округ та ознайомитися з його голосуваннями.



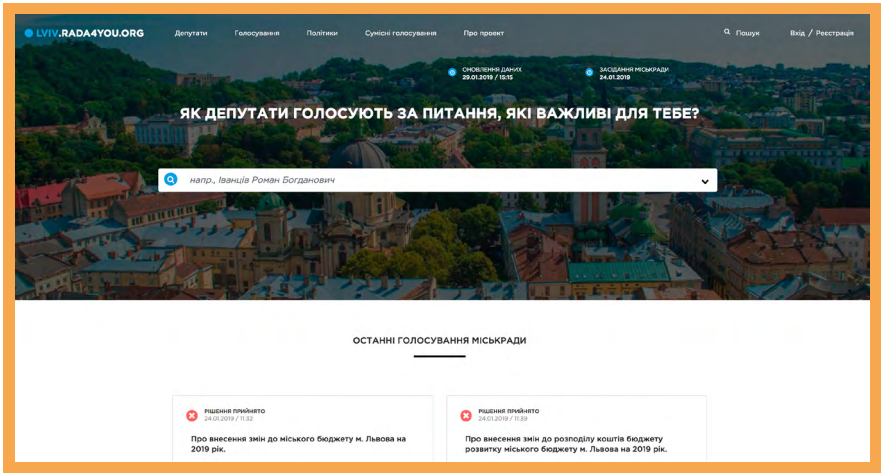
Young Europeans[8] – портал використовує відкриті дані Eurostat для того, щоб створювати демографічні скетчі з інформацією про європейську молодь віком від 16 до 29 років. Хороший інструмент для тих, хто бажає моніторити економічні можливості для свого покоління у сусідніх країнах.



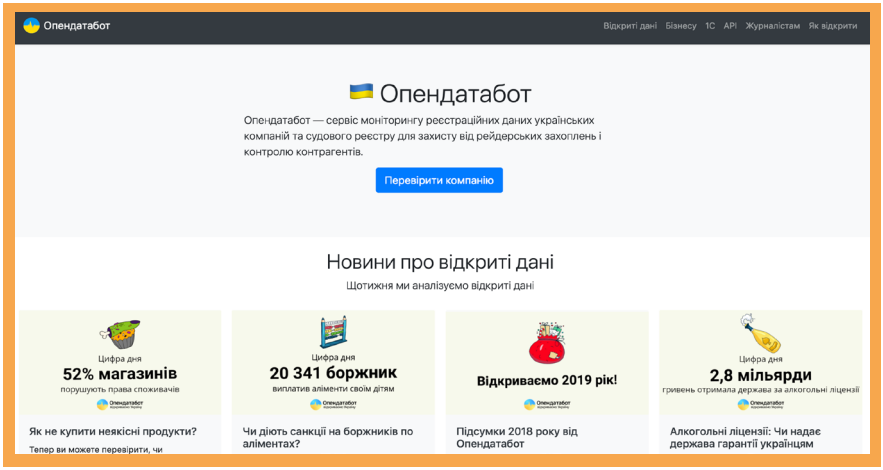
[7] *https://openstates.org/*

[8] *https://ec.europa.eu/eurostat/cache/infographs/youth/index_en.html*

Local Rada4You[9] – система сервісів, які допомагають контролювати депутатів місцевих рад, не відходячи від комп'ютера. Портал опрацьовує дані системи "Віче", якою користуються міські обранці для голосувань. За допомогою ресурсу, можна дізнатися, скільки засідань депутат прогуляв, які рішення підтримав, з якими колегами голосував в унісон.



Опендатабот[10] — платформа моніторингу реєстраційних даних українських компаній та судового реєстру для захисту від рейдерських захоплень і контролю контрагентів. Сервіс працює в популярних мессенджерах — Telegram, Facebook Messenger, Skype, Viber.



[9] *https://lviv.rada4you.org/*

[10] *https://opendatabot.ua/*

Законодавство

9 квітня 2015 року було внесено зміни[11] до ЗУ "Про доступ до публічної інформації" у частині публічної інформації в форматі відкритих даних. Зокрема, було визначено, що "**публічна інформація у формі відкритих даних** – це публічна інформація у форматі, що дозволяє її автоматизоване оброблення електронними засобами, *вільний* та *безоплатний* доступ до неї, а також її *подальше використання*".

21 жовтня 2015 року було прийнято[12] Постанову Кабінету міністрів України №835 про публікацію наборів даних у форматі відкритих даних. В документі було визначено переліки наборів даних, які підлягають обов'язковому оприлюдненню, а також формати даних у яких має відбуватися публікація.

21 листопада 2018 року Кабінет Міністрів України схвалив[13] **Дорожню карту відкритих даних** на 2018-2020 роки. Це вже третя Дорожня карта відкритих даних, яку приймають в Україні. Перша була укладена у 2016 році, коли наша держава приєдналась до Міжнародної хартії відкритих даних[14] та взяла на себе зобов'язання виконувати принципи хартії: перш за все, робити урядові дані **відкритими за замовчуванням**. Дорожня карта має на меті не лише покращити якість, оперативність та доступність відкритих даних, але й прагне популяризувати використання інформації серед журналістів та громадських організацій. Документ також має на меті сприяти створенню місцевих програм розвитку відкритих даних.

На сьогодні українське законодавство також встановлює певні **формати** відкритих даних у яких мають оприлюднюватися відповідні набори інформації. Усі вони повинні забезпечувати можливість автоматизованого оброблення даних *машинним способом*. Варто розуміти, що інформація, яку неможливо опрацювати, втрачає свою цінність для користувачів.

Які формати використовувати для оприлюднення наборів даних?

Тип даних	Формат даних
Текстові дані	TXT, ODT, (X) HTML
Структуровані дані	CSV, JSON, XML, RDF, ODS, YAML
Графічні дані	GIF, JPG (JPEG), PNG
Відеодані	MPEG, MKV, AVI, FLV, MKS, MK3D
Аудіодані	MP3, WAV, MKA
Архів даних	ZIP, 7z

[11] *http://zakon3.rada.gov.ua/laws/show/319-19*

[12] *https://www.kmu.gov.ua/ua/npas/248573101*

[13] *https://zakon.rada.gov.ua/laws/show/900-2018-%D1%80*

[14] *https://opendatacharter.net*

П’ять зірок відкритих даних

Іноді формат оприлюднення даних викликає багато запитань і як наслідок – труднощів. Тож для кращого розуміння важливості коректного оприлюднення відкритих даних давайте звернемося до відомої класифікації **“5 Stars Open Data”**[15], що була розроблена одним із творців Всесвітньої павутини (і ще багатьох речей) **Тімом Бернерсом-Лі**. У даному рейтингу якість та рівень відкритості даних визначається кількістю зірок від 1 до 5 (чим більша цифра – тим краще).

Одна зірка[16] – Ваша інформація доступна в мережі Інтернет у **будь-якому форматі**, але під **відкритою ліцензією**. У цю категорію потрапляють файли у форматі PDF, у тому числі скановані копії документів. Вашу інформацію можна переглянути, роздрукувати та поширити, але опрацювати її без додаткових маніпуляцій (оцифрування) неможливо.

Дві зірки[17] – Ваші дані оприлюднені у структурованому вигляді, проте формат даних не є відкритим (наприклад, XLSX). Багато користувачів для отримання даних залежать від комерційного програмного забезпечення. Ваші дані можна обробляти автоматично, їх можна експортувати в інший формат, проте вони

[15] *https://5stardata.info*

[16] *https://5stardata.info/ru/examples/gtd-1.pdf*

[17] *https://5stardata.info/ru/examples/gtd-2.xls*

все ще містять зайві елементи оформлення, навігації, а значить також потребують додаткових дій для аналізу.

Три зірки[18] – Ваша інформація доступна у відомих та добре описаних відкритих структурованих форматах (наприклад, CSV, JSON, XML, YAML). Користувачі можуть користуватися даними будь-яким чином та без необхідності використання комерційного програмного забезпечення. Проте з іншої сторони, Ваша інформація все ще не є даними, що по-справжньому інтегровані у веб (in the Web)[19].

Чотири зірки[20] – Ви використовуєте **стандарти W3C**[21] (зокрема, RDF та SPARQL), Ваші дані мають постійне посилання. Користувачі можуть отримати первинні набори відкритих даних у вигляді файлів (довідники, списки, таблиці у відкритому форматі, архів документів тощо) або через запит до API за вказаними параметрами. Це дає змогу отримувати тільки потрібну інформацію.

Якщо у Вас є API – він має бути добре описаний, а доступ до нього може бути анонімний без обмежень або з реєстрацією, за вказаним ідентифікатором, лімітами на кількість одночасних запитів тощо.

П'ять зірок[22] – Ваші набори відкритих даних пов'язані між собою (**linked data**). Вони мають спільні довідники, класифікатори, ідентифікатори, посилання між документами та іншими елементами тощо. Дані являють собою семантичну мережу, що постійно оновлюється й змінюється відповідно до сучасних запитів. Від мережевого ефекту виграють і користувач, і публікатор.

Таким чином, відкритість та якість даних залежить не лише від форматів у яких здійснюється публікація, але й від **способів доступу** до інформації та **кількості додаткових дій**, які необхідні для її отримання, збереження та використання.

Запитання та відповіді

Які дані повинні бути першочергово опубліковані?

Опублікувати дані, які найдешевше і найлегше довести до формату відкритих даних, – не кращий підхід до оприлюднення відкритих даних. Влада повинна добре подумати про пріоритезацію: які дані вона хоче оприлюднити першими і скільки це забере часу. Найкращий спосіб вирішити це питання – провести аудит даних (про це детальніше в окремому розділі).

[18] *https://5stardata.info/ru/examples/gtd-3.csv*

[19] *https://webofdata.wordpress.com/2010/03/01/data-and-the-web-choices*

[20] *https://5stardata.info/ru/examples/gtd-4*

[21] *https://www.w3.org/standards*

[22] *https://5stardata.info/ru/examples/gtd-5*

Які правила використання відкритих даних?

Правила використання переважно містять застереження та згадку про обмежену відповідальність щодо даних, а також наступну інформацію:

- Без оплати та вимоги реєстрації.
- Без обмежень щодо використання.
- Без ліцензійних обмежень (всі дані повинні бути доступні).
- Без обов'язкового посилання (за світовою практикою користувачі даних в своїх продуктах посилаються на першоджерело даних).

Як бути з ліцензією?

Не повинно бути ніяких ліцензій, які створюють бар'єр для повторного використання наборів даних.

Авторське право та інші закони в усьому світі автоматично поширюють захист авторських прав на твори авторства та бази даних, незалежно від того, хоче автор чи творець ці права чи ні. Для уникнення цього, можна використовувати формати Creative Commons Zero[23] (**CC0**) чи Open Data Commons Public Domain Dedication and License[24] (**PDDL**).

Наприклад, використовуючи CC0, Ви передаєте набори даних у **суспільне надбання**, відмовляючись від авторських та суміжних прав. Ваші дані можна копіювати, змінювати, поширювати та використовувати, **навіть у комерційних цілях**, не запитуючи дозволу.

На відміну від ліцензій, що також були розроблені організацією Creative Commons, формат CC0 **не передбачає** обов'язкової атрибуції (**зазначення авторства**). Так само, як і будь-що із суспільного надбання, користувачі можуть використовувати та адаптувати дані згідно свої потреб. Варто відзначити, що хоч і атрибуція не є обов'язковою з точки зору юридичних вимог CC0, проте це не означає, що Ви не можете рекомендувати зазначати першоджерело даних відповідно до міжнародної практики та етичних міркувань.

Якщо Ви вирішили обрати формат CC0, Вам потрібно за допомогою простого онлайн інструменту згенерувати[25] HTML-код із вбудованими метаданими для відповідного маркування даних.

Яким має бути портал для публікації відкритих даних?

Якщо у Вас немає можливості створити окремий портал відкритих даних міської ради, ось перелік способів як це зробити без нього:

[23] *http://creativecommons.org/about/cc0*

[24] *https://opendatacommons.org/licenses/pddl*

[25] *https://creativecommons.org/choose/zero*

Перш за все, є можливість публікувати відкриті дані на Єдиному порталі відкритих даних України *data.gov.ua*.

Ви можете створити окремий розділ "Відкриті дані" на локальному сайті і там публікувати дані, наприклад, у форматі CSV.

Можна зареєструватись на GitHub[26] і там ділитися даними із громадянами, проте на офіційному сайті потрібно обов'язково зробити відповідне посилання.

[26] *https://github.com/*

ВПОРЯДКОВАНІ ДАНІ

Що таке впорядковані дані

Впорядковані дані *(англ. Tidy Data) – це добре структуровані дані, які не потребують додаткової очистки та маніпуляцій для їхньої обробки машиночитним способом. Такі набори даних організовані так, що кожна змінна є стовпчиком, а кожне спостереження є рядком.*

Хедлі Вікхем у "Журналі статичного програмного забезпечення" перефразовує[27] Льва Толстого і зазначає: "Всі структуровані набори даних схожі, натомість кожен брудний набір брудний по-своєму". Автор також наводить досить популярну статистику: "80% аналізу даних – це час витрачений на їхню підготовку". Саме тому важливо забезпечити не лише публікацію інформації, але й високу якість та структурованість даних.

Безумовно усі набори даних відрізняються, бо несуть різну інформацію. Проте відомий дата-вчений **Джефф Лік** у своїй книзі[28] "Елементи аналітичного стилю даних" підсумовує чотири головні характеристики будь-яких чвпорядкованих даних:

- Кожна змінна (***variable***), яку ви вимірюєте, повинна бути в **одному стовпці**.
- Кожне окреме спостереження (***observation***) цієї змінної – в **окремому рядку**.
- Для **кожного "виду"** змінної має бути одна таблиця.
- Якщо у вас є **декілька таблиць** – вони повинні включати стовпець (**ідентифікатор**) у таблиці, завдяки якому їх можна поєднати.

У простому розумінні, значна частина даних – це таблиці. ***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.

[27] *https://vita.had.co.nz/papers/tidy-data.pdf*

[28] *http://worldpece.org/sites/default/files/datastyle.pdf*

Структура даних

Цілком зрозуміло, що в реальному світі, знайти якісний набір даних можна не завжди, а іноді навіть дуже рідко. Тому для правильного створення та оприлюднення даних варто не лише уникати типових помилок, але й дбати про структуру даних. Давайте розглянемо найчастіші проблеми.

Порушення структури рядків та стовпців (об'єднані комірки)

Таблиця з об'єднаними комірками *(Таблиця 1.)* точно не є відкритими даними. І без попередньої "чистки" фактично не є даними як такими. У такому наборі неможливо навіть відфільтрувати інформацію, не кажучи вже про машиночитну обробку.

Таблиця 1.

	чисельність наявного населення		
	всього	у тому числі	
		міське	сільське
1990	51838,5	34869,2	16969,3
1991	51944,4	35085,2	16859,2
1992	52056,6	35296,9	16759,7
1993	52244,1	35471,0	16773,1
1994	52114,4	35400,7	16713,7
1995	51728,4	35118,8	16609,6
1996	51297,1	34767,9	16529,2
1997	50818,4	34387,5	16430,9
1998	50370,8	34048,2	16322,6
1999	49918,1	33702,1	16216,0
2000	49429,8	33338,6	16091,2

Звісна річ, що дані для презентації або звіту цілком можуть мати об'єднані комірки та порушувати табличну структуру, проте такі набори **не можна публікувати в якості відкритих**. Крім того, варто подумати про доцільність існування у наборі змінних, які можна отримати за допомогою **простих арифметичних дій** (наприклад, загальна чисельність населення) *(Таблиця 2.)*.

Таблиця 2.

рік	тип населення	чисельність
1990	міське	34869.2
1990	сільське	16969.3
1991	міське	35085.2
1991	сільське	16859.2
1992	міське	35296.9
1992	сільське	16759.7
1993	міське	35471
1993	сільське	16773.1
1994	міське	35400.7
1994	сільське	16713.7
1995	міське	35118.8
1995	сільське	16609.6

Правильно структурована таблиця придатна для аналізу даних. А за потреби, завжди можна легко змінити її структуру.

В заголовках стовпців знаходяться значення, а не назви змінних

Дана таблиця може видатися досить зручною для читачів, адже непогано демонструє динаміку середньої заробітної плати у вибраних містах України. Людині "на око" зручно порівнювати цифри за роками та містами. Проте таблиця мало придатна для машинної обробки: її неможливо сортувати, складно фільтрувати. А що робити, якщо кількість років буде більшою?

Таблиця 3.

місто	2016	2017	2018
Київ	8585	11134	13547
Львів	5858	7521	11855
Одеса	6967	8469	11909
Харків	6365	8311	11546

Насправді у цьому наборі даних три змінних: "*місто*", "*рік*" та "*середня зарплата*" *(Таблиця 4.)*. Такий структурований набір даних може гірше сприйматися людиною, проте його значно легше аналізувати машиночитним способом.

Таблиця 4.

місто	рік	середня зарплата
Київ	2016	8585
Київ	2017	11134
Київ	2018	13547
Львів	2016	5858
Львів	2017	7521
Львів	2018	11855
Одеса	2016	6967
Одеса	2017	8469
Одеса	2018	11909
Харків	2016	6365
Харків	2017	8311
Харків	2018	11546

Кілька змінних зберігаються в одному стовпці.

Дана таблиця дещо схожа на попередню, проте у даному випадку змінні "рік" та "професія" взагалі зберігаються в одному стовпці. Натомість має бути навпаки. Адже для кожної змінної має бути окремий стовпець.

Таблиця 5.

місто	2017_аудит	2017_дизайн	2018_аудит	2018_дизайн
Київ	9169	11141	11892	14449
Львів	8087	9889	10383	12697
Харків	8253	9208	10033	11194
Одеса	7239	10319	9453	13474

Структурувавши таблицю, отримаємо чотири змінні: "місто", "рік", "професія" та "середня зарплата". Таким чином, ми можемо здійснювати аналіз даних одразу за декількома параметрами.

Таблиця 6.

місто	рік	професія	середня зарплата
Київ	2017	аудит	9169
Львів	2017	аудит	8087
Харків	2017	аудит	8253
Одеса	2017	аудит	7239
Київ	2017	дизайн	11141
Львів	2017	дизайн	9889
Харків	2017	дизайн	9208
Одеса	2017	дизайн	10319
Київ	2018	аудит	11892
Львів	2018	аудит	10383
Харків	2018	аудит	10033
Одеса	2018	аудит	9453
Київ	2018	дизайн	14449
Львів	2018	дизайн	12697
Харків	2018	дизайн	11194
Одеса	2018	дизайн	13474

Змінні у таблиці не є уніфікованими, "брудні дані"

Дуже часто, навіть маючи правильно структуровану таблицю, ми не можемо працювати із даними, через те, що вони є "брудними". Ми не зможемо сортувати дані за змінною "місто" чи "підприємство", оскільки кожен раз **назви пишуться по-різному**. У таблиці є часовий вимір, однак жодних маніпуляцій із датами ми зробити не зможемо, оскільки вони **мають різний формат**. Порахувати прибуток ми теж не зможемо, оскільки дані у цьому стовпці більше

схожі на набір символів, ніж на числа.

Таблиця 7.

місто	підприємство	дата	прибуток
Київ	Комунсервіс	2 червня 2017	14439934
Киев	ДП "Комунсервіс"	2016-04-01	1 млн122 тис 14
м. Київ	"Комун-сервіс"	23.01.2018	2 000 000
Львив	Львівкартон	1 лист 2016 р.	1,567,890
Львівв	Львів картон	23.08.2019	768431
Харків	Льоова арена	2015-03-11	+2300943
Харьков	Льодова арена	14/2/2012	5487643

Натомість впорядковані дані можна фільтрувати, сортувати, аналізувати, візуалізувати, створювати на їхній основі сервіси *(Таблиця 8.)*. За потреби такий набір даних можна легко доповнити новою інформацією.

Таблиця 8.

місто	підприємство	дата	прибуток
Київ	ДП "Комунсервіс"	2017-06-02	14439934
Київ	ДП "Комунсервіс"	2016-04-01	1122014
Київ	ДП "Комунсервіс"	2018-01-23	2000000
Львів	ДП "Львівкартон"	2016-11-01	1567890
Львів	ДП "Львівкартон"	2019-08-23	768431
Харків	ДП "Льодова арена"	2015-03-11	2300943
Харків	ДП "Льодова арена"	2012-02-14	5487643

Поради щодо уникнення типових помилок

- В жодному разі не використовуєте **об'єднані комірки**.
- Заповнюйте всі комірки в таблиці, навіть коли дані відсутні — внесіть в них запис "**NA**" (Not Available). Використовуйте завжди один і той самий запис для пропущених значень. Не користуйтеся значеннями "0", "-", "999", "..." тощо.

- Вставляйте лише **одне значення** в комірку. Наприклад, у комірці може бути записана відстань – "19 км". Краще написати просто "19", а одиниці виміру винести в назву колонки, наприклад "відстань_км". Проте найкращим варіантом буде, якщо Ви назвете колонку "відстань", а одиниці виміру винести в словник (структуру) набору даних.
- У межах **однієї змінної** вживайте **один тип даних.** Тобто, якщо в певному стовпчику записані дати, у ньому не повинно бути текстових даних чи чисел.
- Якщо Ви працюєте з даними в Excel – вірно обирайте **формати стовпців** (текст, число, дата).
- Записуйте усі дати в **одному форматі**. Використовуйте один загальний формат для всіх дат. Бажано використовувати стандарт ISO 8601, тобто РРРР-ММ-ДД (рік-місяць-день). Наприклад, "2019-02-21".
- Використовуйте **один простий формат для чисел** – не слід вдаватися до зайвого форматування, наприклад розділення великого числа комами чи пробілами для його кращого візуального сприйняття. Використовуйте крапку як **десятковий розділювач**. Наприклад, "108.7" замість "108,7".
- Коли потрібно зберегти **провідні нулі** (які йдуть попереду числа, наприклад, при використанні кодів бюджету "02509000000"), форматуйте комірки з числами як текст.
- Ваші дані не мають містити в комірках результатів **підрахунків чи формул**.
- **Один стовпчик** у таблиці – **одна змінна**. Поширена помилка – використовувати значення змінних (назви областей чи років) в якості назви стовпців даних в таблиці.
- Не використовуйте **перенос рядка** в текстових комірках.
- Не використовуєте **латинські літери замість кириличних**, і навпаки. Наприклад, назва міста "Сарни" (українськими літерами) та "Сарни" (з першою латинською літерою) виглядають однаково, проте на практиці є різними значеннями.
- Будьте уважними **з пробілами**. На початку і в кінці комірки з даними не має бути пробілів. Не можна ставити два пробіли підряд. Наприклад, назва міста "Горішні Плавні" буде відрізнятися від назви "Горішні Плавні" із зайвими пробілами.
- **Порожня комірка** (дані відсутні) буде відрізняється від ніби порожньої комірками, але з пробілами.
- Не потрібно писати текстові значення чи назви стовпців **ВЕЛИКИМИ ЛІТЕРАМИ** (крім абревіатур). Під час обробки даних машиною це може сприйматись як різні значення.

- Якщо Ви використовуєте стовпець з універсальними **ідентифікаторами** (для можливості поєднання даних) – вони мають бути записані відповідно до однієї системи. Наприклад, "1000", "1001", "1002", а не "kod_1", "id_2", "kod3".
- **Завжди використовуйте кодування "UTF-8"**.
- Перевірте чи немає у наборі даних **дублікатів**.
- Називайте Ваші файли лише **латиницею**. Наприклад, "budget.csv"
- Не варто використовувати пробіли в заголовках стовпців або назвах файлів. Замість них використовуйте **підкреслення** "_". Наприклад, "budget_2019.csv"

Колір та шрифт у наборі даних

*Колір та шрифт комірки **не є даними**. Ці функції форматування можуть бути корисними для візуального представлення чи сприйняття даних, проте вони не несуть ніякого значення під час автоматичної обробки даних. Тому якщо інформація, закодована за допомогою кольору чи шрифта, є справді важливою, обов'язково додавайте її **окремою змінною**.*

Як зробити дані чистими?

Як ми вже розглянули вище, важливою цінністю даних, окрім їхнього вільного використання, є їхня якість. Проте чисті дані – це результат наполегливої праці. Для того, щоб її полегшити та/чи прискорити – можна скористатися як вбудованими функціями Excel чи Google Spreadsheets, так і спеціальними інструментами для перевірки/очищення даних, наприклад Data Proofer та OpenRefine.

Excel

- Для багатьох випадків чистки помилок чи одруківок у даних досить використання команди **пошуку і заміни** (англ. Find & Replace, комбінація клавіш Ctrl + H). Таким чином можна замінити лапки чи розділювач, прибрати зайві пробіли, помилкові символи, уніфікувати текстові значення, якщо їх не дуже багато тощо.
- Для об'єднання вмісту двох або більше комірок можна використати функцію "**CONCATENATE.**" Наприклад, =CONCATENATE(A2," ",B2) об'єднає вміст комірок A2 та B2, розділивши їх пробілом.
- Для розділення комірок до різних стовпців, що використовують певний символ-роздільник, наприклад, крапку чи крапку з комою, можна використати функцію "**текст за стовпцями**" (англ. Text to Columns), що знаходиться на вкладці "Дані".

- Для видалення даних, що повторюються можна також скористатися базовою функцією "**видалення дублікатів**" (англ. Remove Duplicates), що знаходиться на вкладці "Дані".
- Для очищення комірок від зайвих пробілів можна використати функцію "**TRIM**".

Більш детальні інструкції щодо можливостей структурування та очистки даних в Excel можна отримати з посібника[29] "Відкриті дані: формати і правила створення", який підготували *Texty.org.ua*.

Data Proofer

Для виявлення проблем у даних можна скористатися безкоштовною програмою Dataproofer[30], яка застосовує до кожного набору даних 14 різних тестів. Вона допоможе перевірити набори даних у табличних форматах: xls(x), CSV, TSV та PSV.

За допомогою програми можна:

- Перевірити чи всі значення в стовпцях є числами чи текстом.
- Виявити наявність дублікатів чи пустих комірок.
- Побачити потенційну втрату даних: коли, наприклад, загальна кількість рядків у таблиці становить 65 536 (обмеження попередніх версій Excel) чи загальна кількість символів у комірці – 255 (деякі обмеження під час експорту з баз даних)
- Перевірити недійсні чи пусті значення широти та довготи.
- Провести тексти, які перевірять відхилення від середнього значення та/чи медіани стовпця.

OpenRefine

Для очищення наборів даних чи перетворення їх в інші формати можна використовувати спеціальний інструмент – Open Refine[31]. Ця програма виконується локально і працює всередині веб-браузера, який запускається автоматично. Open Refine підтримує імпорт даних з форматів CSV, *SV, xls(x), JSON, XML тощо.

Програма дозволяє:

- Виявляти та виправляти помилки в даних, зокрема — знаходити різні варіанти написання назв і дат та привести їх до єдиного вигляду.

[29] *http://texty.org.ua/pdf/data2017.pdf*

[30] *http://dataproofer.org*

[31] *http://openrefine.org*

- Знаходити та видаляти зайві символи та комбінації символів у даних, в тому числі за допомоги регулярних виразів.
- Розбивати стовпчики за певним роздільником або ж навпаки об'єднувати дані із певним роздільником.
- Фільтрувати дані за кількома показниками.
- Виділяти/видаляти стовпчики, змінювати структуру документу.
- Здійснювати базовий аналіз даних.
- Конвертувати дані в різні формати.
- Надсилати запити до різних API (наприклад, геокодувати адреси).
- Зберігати всі трансформації даних в окремому проекті та застосовувати їх до інших документів.

R

За допомогою програмного середовища R також можна здійснювати обробку, структурування чи поверхневу чистику даних. Для багатьох операцій Вам буде достатньо навіть простих навичок програмування. Ось деякий перелік бібліотек, які можуть допомогти при структуруванні/очищенні даних:

- **stringr** – допомагає краще працювати з текстовими рядками. Можна дізнатися довжину символів в комірці, здійснити необхідні заміни чи прибрати зайві пробіли.
- **lubridate** – набір функцій для роботи з датами та часом. Бібліотека, наприклад, може визначити день тижня за датою чи порахувати проміжок між двома датами.
- **tidyr** – робота із "брудними" даними, зокрема їхнє структурування. Також за допомогою бібліотеки також можна розділити чи об'єднати комірки.

Словник набору даних

Дуже часто із Вашими даними можуть працювати одні й ті ж самі люди, які вже "звикли" до них та "вивчили". Вони добре знають, яка інформація їм потрібна та яким чином її краще обробити. Але як щодо користувачів, які вперше почали працювати із Вашими даними? Чи може їм щось сказати цей набір без відповідного опису? Якщо й так, то скільки часу займе розібратися у значеннях кожної зі змінних? А таких наборів може бути десятки, а то й сотні.

Таким чином, для опису даних потрібно створити словник (структуру) набору даних — окрему таблицю метаданих (даних про дані). У ній необхідно описати значення кожної зі змінних, їхній тип (текст, число, дата тощо), одиниці виміру, мінімальні і максимальні значення, якщо такі можуть бути. До словника набору даних також варто включити усі необхідні примітки, уточнення, розшифровки тощо.

Наприклад, ми маємо набір даних із усіма спортивними секціями, які функціонують у місті Вінниця *(Таблиця 9.)*. У таблиці є 6 змінних: "id", "sport", "organization", "date", "address" та "contact". Їх усіх потрібно пояснити.

Таблиця 9.

id	sport	organization	date	address	contact
1	Баскетбол	МСДЮСШОР	2015-05-02	вул. Пирогова, 4	69-71-49
2	Бокс	МКДЮСШ "Вінниця"	2017-02-23	вул. Замостяна, 16	55-68-67
3	Боротьба вільна	МДЮСШ №5	2017-01-24	вул. Хлібна, 1	56-20-17
4	Боротьба самбо	МДЮСШ №5	2018-09-25	вул. Хлібна, 1	56-20-17
5	Важка атлетика	МДЮСШ №5	2011-02-26	вул. Хлібна, 1	56-20-17
6	Велоспорт	МДЮСШ № 3	2017-02-11	вул. Театральна, 24	67-32-85
7	Веслувальний слалом	МДЮСШ № 2	2017-06-14	узвіз Бузький, 33	67-36-86
8	Веслування на б\к	МДЮСШ № 2	2018-03-01	узвіз Бузький, 33	67-36-86
9	Волейбол	МДЮСШ № 3	2010-03-02	вул. Театральна, 24	67-32-85
10	Гандбол	МДЮСШ № 3	2017-03-13	вул. Театральна, 24	67-32-85
11	Гімнастика художня	МДЮСШ № 1	2018-03-04	вул. Хлібна, 1	67-12-47
12	Дзюдо	МДЮСШ №5	2017-10-25	вул. Хлібна, 1	56-20-17
13	Легка атлетика	МДЮСШ № 1	2019-03-06	вул. Хлібна, 1	67-12-47
14	Пауерліфтинг	МДЮСШ №5	2011-03-07	вул. Хлібна, 1	56-20-17
15	Радіоспорт	МДЮСШ № 2	2006-03-08	узвіз Бузький, 33	67-36-86

Пояснюємо кожну зі змінних в окремій таблиці – словнику (структурі) набору даних. Вона також має бути добре структурована.

Таблиця 10.

variable	description	type
id	Унікальний ідентифікатор виду спорту.	integer
sport	Назва виду спорту (секції).	string
organization	Назва закладу, де відбуваються тренування.	string
date	Дата створення секції, у форматі рррр-мм-дд.	date
address	Адреса організації. Усі заклади знаходяться у м. Вінниця.	string
contact	Контактний телефон організації (для довідок). Код міста – (0432).	string

Тепер користувачі зможуть отримати чіткі пояснення щодо даних з якими вони працюють. Адже іноді інформація, яка публікується, може містити специфічні терміни, позначення чи класифікації, які відомі лише розпоряднику. Пам'ятайте про це!

ФОРМАТИ ДАНИХ

Формат даних *(формат файлів) – це інформація за допомогою якої користувачі (як і операційна система) можуть швидко ідентифікувати вміст даних без необхідності зчитування вмісту всього файлу.*

Саме тому до вибору формату даних треба підходити відповідально. Під час визначення формату для оприлюднення набору даних необхідно зважати саме на відповідність типу даних файловому формату. Найбільш поширеною помилкою, яка виникає під час публікації наборів даних, є невідповідність файлового формату типу даних, що у ньому міститься. Зокрема, некоректною є публікація таблиць (структурованих даних) у форматах DOC(X) чи PDF, призначених для текстових даних, або у форматах JPG чи PNG, призначених для графічних даних.

Крім того, створювати набори даних потрібно у так званих відкритих файлових форматах, тобто таких, що не залежать від платформи та доступні без обмежень.

До відкритих форматів, зокрема, належать формати, CSV, JSON та XML. Давайте їх коротко розглянемо детальніше.

.csv

CSV *(від англ. comma-separated values, "значення, розділені комою")* – простий текстовий формат, призначений для представлення табличних даних.

- Кожен рядок файлу – це один рядок таблиці.
- Рядки розділені знаком нового рядка \n.
- Роздільником (delimiter) значень стовпців найчастіше є символ коми ",". Проте дуже часто на практиці використовують й інші роздільники, зокрема крапка з комою (semicolon) чи табуляція (tap).
- Значення, що містять так звані зарезервовані символи (лапки, кома, крапка з комою, новий рядок) охоплюються подвійними лапками ("). Якщо у значенні зустрічаються лапки – вони представляються у файлі у вигляді двох лапок поспіль.

Наприклад, csv-файл із інформацією про маршрути, їхню довжину, вартість проїзду та перевізника буде виглядати так:

```
Маршрут,Довжина,Вартість,Перевізник
1,24.97,12,ФОП Попович
2,41.45,21,ФОП Баштовий
3,19.48,9,"ТОВ""АТП-1234"""
4,28.63,14,ФОП Попович
5,34.13,17,"ТОВ""Атлант"""
6,39.62,20,ФОП Величко
7,21.31,10,ФОП Попович
```

Даний набір даних можна легко імпортувати, наприклад, до Excel (Вкладка "Дані" –> "З тексту") та отримати інформацію у вигляді таблиці:

Маршрут	Довжина	Вартість	Перевізник
1	24.97	12	ФОП Попович
2	41.45	21	ФОП Баштовий
3	19.48	9	ТОВ "АТП-1234"
4	28.63	14	ФОП Попович
5	34.13	17	ТОВ "Атлант"
6	39.62	20	ФОП Величко
7	21.31	10	ФОП Попович

Загалом .csv є дуже простим та відносно компактним форматом даних. Він може доволі легко сприйматися людиною, його можна відкрити величезною кількістю програм. Проте даний формат також має ряд недоліків:

- Не підтримує ієрархію даних.
- Не підтримує зв'язок між даними.
- Придатний лише для таблиць.
- Заголовки в таблиці не є обов'язковими.

Увага!

1. *Excel за замовчуванням використовує кодування Вашої операційної системи.*
2. *В Україні в якості десяткового розділювача чисел здебільшого використовується кома ",".*

Таким чином, Ваш Excel зберігатиме csv-файли з розділювачем ";" і комами замість крапок в не цілих числах. Для уникнення цієї проблеми можна скористатися Google Spreadsheets, LibreOffice Calc чи OpenRefine.

Для конвертування .csv до інших форматів можна скористатися безкоштовним ресурсом Convert CSV[32].

.JSON

JSON *(англ. JavaScript Object Notation)* – текстовий формат обміну даними, що

[32] *http://www.convertcsv.com*

заснований на JavaScript. Якщо Ви маєте великий ієрархічний масив даних – обирайте JSON.

Загалом даний формат:

- компактний та структурований;
- стандартизований (RFC 7159, ECMA 404);
- завдяки ієрархічній структурі дозволяє зменшити розмір файлу з ієрархічними даними;
- за наявності певного досвіду чи спеціальних розширень (для браузера Google Chrome – це JSONView) може легко сприйматися людиною.

Цей формат чудово підходить для передачі даних через API.

Оскільки .json не чутливий до відступів, таблицю про маршрути, яку ми вже розглядали вище, у цьому форматі можна подати так:

```
[
    {
            "Маршрут":1,
            "Довжина":24.97,
            "Вартість":12,
            "Перевізник":"ФОП Попович"
    },
    {
            "Маршрут":2,
            "Довжина":41.45,
            "Вартість":21,
            "Перевізник":"ФОП Баштовий"
    },
    {
            "Маршрут":3,
            "Довжина":19.48,
            "Вартість":9,
            "Перевізник":"ТОВ \"АТП-1234\""
    }
]
```

Попри те, що формат JSON відносно легко можуть читати та редагувати люди, його варто створювати машинним шляхом. Крім того, перед публікацією (особливо, якщо Ви редагували файл вручну) JSON-файли можна перевірити за допомогою спеціальних сервісів: **JSON Formatter**[33] або **JSON Editor Online**[34]. Крім того, дані ресурси підтримують і редагування файлів.

[33] *https://jsonformatter.curiousconcept.com*

[34] *http://jsoneditoronline.org*

Важливо! *Ваш JSON має бути у кодуванні UTF-8. У іншому випадку це може спотворити символи, зокрема кириличні і користувачі замість даних побачать:*

```
{
dep: [
{
name: "?????? ??? ?? ???? ????? ???? ???",
id: "279",
kom: [
{
kom_name: "?????? ? ?? ???? ???? ??? ??? ???? ?? ??? ???????",
kom_id: "19",
reg: [
{
reg_date: "05.12.2014 12:00:09",
result_reg: "³???????"
},
{
reg_date: "10.12.2014 15:00:59",
result_reg: "??? ??????"
},
```

.XML

XML (англ. eXtensible Markup Language) – ієрархічний формат даних, створений ще у 1994 році та рекомендований Консорціумом Всесвітньої павутини (W3C). У ньому дані організовані в об'єкти, які можуть містити інші об'єкти. Однією з головних переваг XML є його гнучкість.

У простому розумінні XML – це конкретна граматика із своїм набором правил, яка, проте, не вимагає формальних, фіксованих назв тегів чи параметрів, тому кожен розробник може створювати свою розмітку відповідно до своїх потреб, дотримуючись загальних правил синтаксису. По суті XML – це мова, яка описує себе і будь-які структури даних.

Беззаперечною перевагою XML є те, що у файлі, крім основних даних можна розміщати метадані (описи, характеристики, реквізити), вкладені файли (наприклад, картинки, стилі тексту), довідники тощо.

Проте, попри ряд значних переваг, XML має і суттєві недоліки: за рахунок повторення тегів та відступів файли цього формату можуть бути значними за розміром. Крім того, у порівнянні з CSV та JSON, даний формат є значно складнішим для парсингу та обробки.

Вже розглянуті нами дані про маршрути та вартість проїзду у форматі XML можуть виглядати так:

```
<?xml version="1.0"?>
<document>
 <row>
  <Маршрут>1</Маршрут>
  <Довжина>24.97</Довжина>
  <Вартість>12</Вартість>
  <Перевізник>ФОП Попович</Перевізник>
 </row>
  <row>
   <Маршрут>2</Маршрут>
   <Довжина>41.45</Довжина>
   <Вартість>21</Вартість>
   <Перевізник>ФОП Попович</Перевізник>
 </row>
  <row>
   <Маршрут>3</Маршрут>
   <Довжина>19.48</Довжина>
   <Вартість>9</Вартість>
   <Перевізник>ТОВ "АТП-1234"</Перевізник>
 </row>
</document>
```

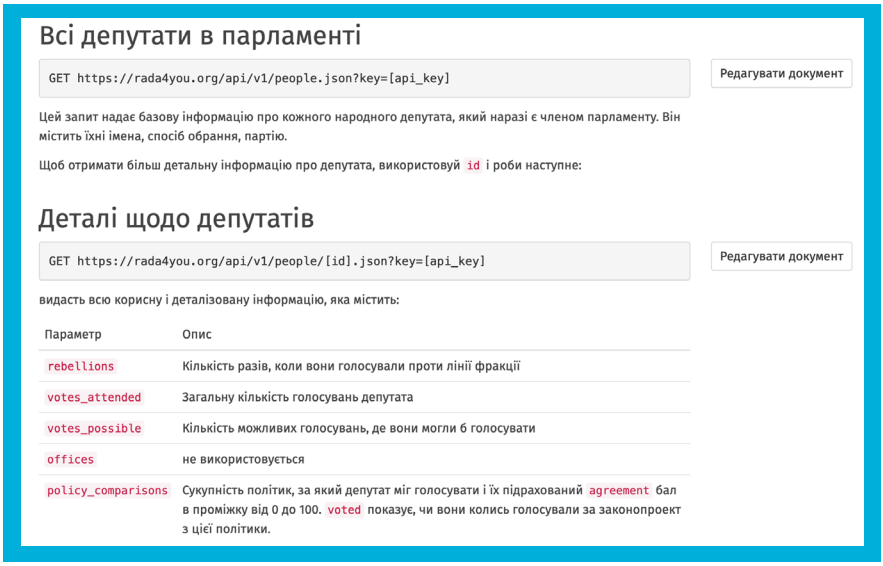
API

API *(англ. Application Programming Interface, "Прикладний програмний інтерфейс")* – набір готових процедур, підпрограм, функцій, посилань чи параметрів, які дозволяють використовувати інформаційні системи для отримання структурованих або неструктурованих наборів даних чи іншої взаємодії. API є зручним та корисним для великих обсягів динамічних даних, оскільки дає змогу отримувати інформацію в автоматичному режимі та/чи в реальному часі. Якщо ваш набір даних містить великий обсяг інформації й часто оновлюється, доступ до нього варто запровадити саме за допомогою API.

Проте варто пам'ятати, що Вам також потрібно надати детальні інструкції щодо використання API, приклади запитів зі всіма можливими параметрами та відповідей на ці запити.

Наприклад, портал "Вони голосують для тебе"[35] пропонує отримати доступ до практично всієї інформації, яка міститься на сайті: інформація про депутатів, їхні голосування, ініціативи, які вони підтримують, голосування, які відбуваються у Верховній Раді. Для цього потрібно зареєструватись (варто ввести email та придумати логін і пароль) і Ви отримаєте спеціальний ключ, який допоможе отримувати актуальні, структуровані та регулярно оновлювані набори даних у форматі JSON.

[35] *https://rada4you.org/help/data*



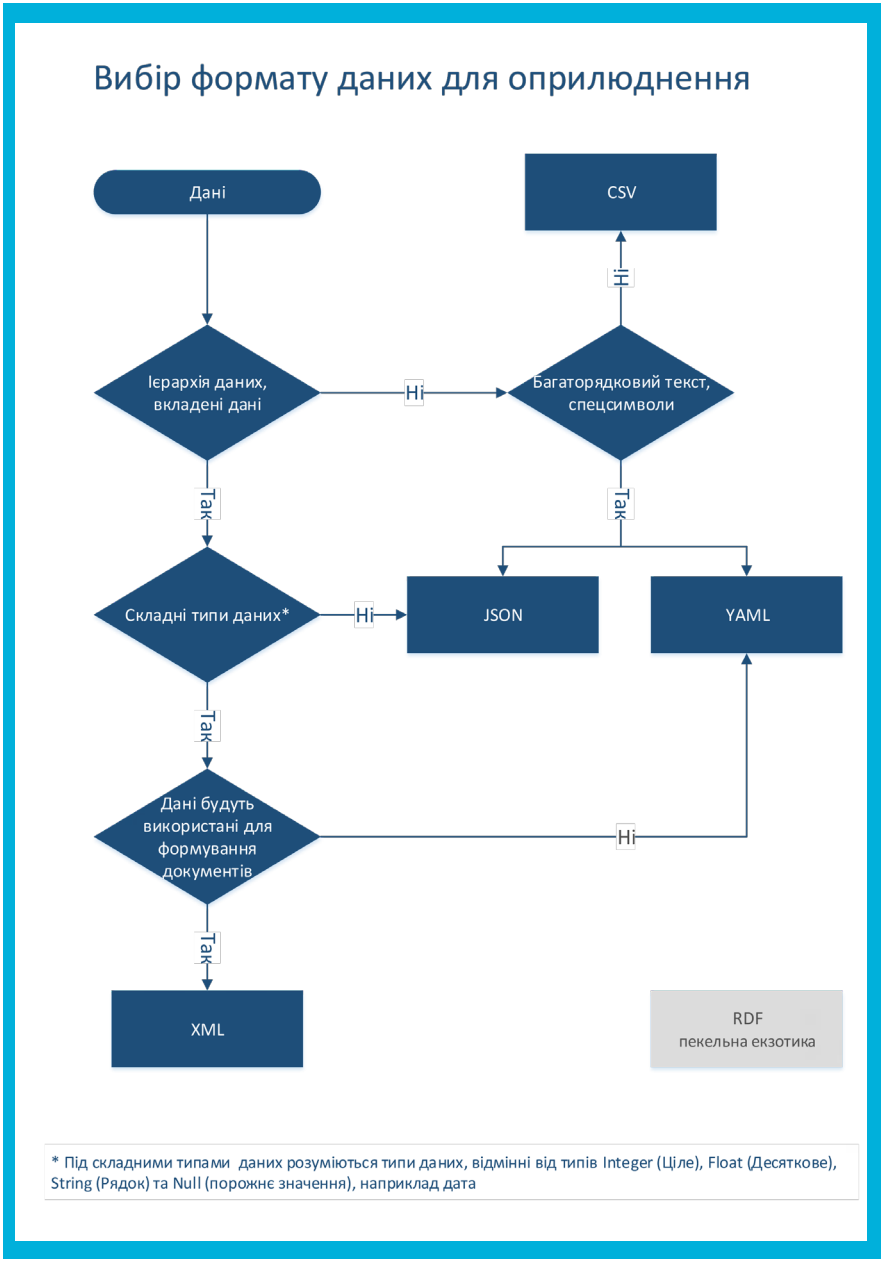
Ваш API має бути добре описаний!

Вибір формату даних

На жаль, неможливо написати ідеальну та уніфіковану інструкцію з вибору формату даних, адже кожен набір даних унікальний. Проте завжди можна скористатися загальними правилами:

1. Якщо у Вас є якісні та структуровані дані (з великою ймовірністю це таблиці Excel), наприклад, із інформацією про комунальні підприємства, перевізників, школи, лікарні тощо – конвертуйте їх у формат CSV та опублікуйте.
2. Якщо потрібно оприлюднити набір даних із ієрархією (наприклад, результати голосувань) обирайте JSON або ж XML.
3. Формат XML варто обирати лише в тому випадку, якщо у Вас уже є робочий, перевірений та грамотно розроблений шаблон структури даних або ж Ви здійснюєте експорт із Вашого робочого середовища. У іншому випадку – краще обирати JSON.
4. Якщо Ваші набори даних регулярно оновлюються – краще розробити API.
5. Незалежно від формату даних, завжди обирайте кодування “UTF-8”.
6. Великі за розміром файли (понад 200 Мб) варто архівувати. Ваші архіви повинні бути у форматі ZIP або 7z.

Для вибору формату даних для оприлюднення також можна скористатися спеціальною схемою:



АУДИТ ДАНИХ

Основні засади аудиту даних

Аудит даних – це комплексне дослідження наявності, стану, форматів, процесів управління й використання даних, а також вироблення на основі отриманої інформації рекомендацій щодо покращення процесів роботи з даними, максимізації їх використання та розкриття потенціалу[36].

Єдиний державний веб-портал відкритих даних підкреслює важливість проведення аудиту даних і зазначає, що дане дослідження рекомендоване для виконання Постановою Кабінету Міністрів України №835. Адже аудит дає змогу розпоряднику дослідити й зрозуміти, у якому стані наразі дані і що потрібно зробити, щоб налагодити процес оприлюднення необхідних наборів у форматі відкритих даних. Аудит допомагає також виявити **дублювання зусиль розпорядників** щодо збору та оприлюднення даних, виявляє проблемні ділянки, які потребують додаткової роботи, що дає змогу краще розподілити ресурси.

В рамках спільної ініціативи Громадянської мережі ОПОРА та TechSoup Poland "Дані міст/Apps4Cities" була розроблена спеціальна методологія аудиту даних за якою п'ять міст-переможців конкурсу "Відкритий виклик" протягом 2018 року інвентаризували свої дані. Аудит даних міст дозволив інвентаризувати інформацію щодо даних, якими розпоряджаються міські ради Ужгорода, Вознесенська, Коломиї, Рівного та Миколаєва[37].

Унікальність запропонованої методології, що вже апробована ініціативою TechSoup Poland "Дані міст", полягає у тому, що вона дозволяє провести аудит без втручання сторонніх осіб у роботу підрозділів міської ради. Адже останні є розпорядниками так званої чутливої інформації: персональних даних, документів для службового користування тощо.

Варто розуміти, що інвентаризація є першим кроком для будь-якого міста на шляху до повноцінного **впровадження політики відкритих даних**. Тож для оцінки вихідних умов варто поставити чіткі цілі та виконати наступні завдання:

1. Зрозуміти, чи існують дані взагалі, у яких форматах доступні: в цифрових чи лише на папері.
2. Оцінити якість даних: вони агреговані чи дезагреговані, на скільки деталізовані, як часто оновлюються тощо.
3. Дослідити правову та технічну відкритість наборів даних.
4. Зрозуміти, чи дійсно ці дані є корисними для всіх існуючих зацікавлених сторін (далі - стейкхолдерів), зокрема представників органів влади самого міста та його мешканців, бізнесу, організацій громадянського суспільства тощо.

[36] *https://data.gov.ua/uploads/files/2018-08-11-104328.578029Part03.pdf*

[37] *https://audyt.danimist.org.ua*

5. Оцінити доцільність того, наскільки легко/складно було б відкрити дані (юридичні, політичні, технічні, інституційні). Це включає в себе структурування даних, додавання метаданих, очищення та перетворення даних у формати, які автоматично обробляються машинами, а також необхідність та зручність анонімності.
6. Продумати інші не зовсім технічні наслідки відкриття даних, наприклад: потенційні втрати доходу; дані, які не можуть бути опубліковані через державну таємницю або з міркувань безпеки тощо.

Оцінка наявності даних та їх якості

Аудит даних – доволі складний та часозатратний процес. Проте його проведення обов'язково зекономить багато сил у майбутньому. Без якісного аудиту – неможливо повноцінно відкрити дані.

Загалом у цьому процесі можна виділити шість етапів:

- **Етап №1.** Внутрішній аудит. Визначення наявності, формату та структури наборів даних кожним із розпорядників інформації міської ради. Варто заповнити Анкету №1.
- **Етап №2.** Проведення тематичного опитування відповідального підрозділу щодо впровадження політики відкритих даних у місті. Варто заповнити Анкету №2.
- **Етап №3.** Зовнішній аудит. Незалежна сторона самостійно здійснює пошук вже опублікованих наборів даних, оцінює їхню якість.
- **Етап №4.** Зацікавлені сторони документують дані, розробляють та представляють рекомендації на основі аудиту.
- **Етап №5.** Впровадження наданих рекомендацій. Ухвалення необхідної нормативно-правової бази. Створення реєстру наборів відкритих даних.
- **Етап №6.** Реалізація міською владою комплексної політики відкритих даних за замовчуванням. Публікація реєстру наборів відкритих даних для користувачів.

Перед початком аудиту обов'язково потрібно визначити:

- Відповідальну особу за проведення аудиту.
- Терміни проведення аудиту.
- Підрозділи/департаменти, де відбуватиметься аудит.

Основні етапи аудиту даних

1. 
Внутрішній аудит. Визначення наявності, формату та структури наборів даних кожним із розпорядників інформації. Варто заповнити Анкету №1.

2. 
Проведення тематичного опитування відповідального підрозділу щодо впровадження політики відкритих. Варто заповнити Анкету №2.

3. 
Зовнішній аудит. Незалежна сторона самостійно здійснює пошук вже опублікованих наборів даних, оцінює їхню якість.

4. 
Зацікавлені сторони документують дані, розробляють та представляють рекомендації на основі аудиту.

5. 
Впровадження наданих рекомендацій. Ухвалення необхідної нормативно-правової бази. Створення реєстру наборів відкритих даних.

6. 
Реалізація міською владою комплексної політики відкритих даних за замовчуванням.

Етап №1

Перш за все дані, якими володіють та розпоряджаються міські ради, варто проаналізовувати на предмет відповідності таким критеріям:

Доступність:

- чи існують дані взагалі?
- чи доступні дані в цифрових форматах?
- якщо дані викладені у відкритий доступ, то де?
- як можна отримати доступ до даних? чи застосовується прикладний програмний інтерфейс API?
- у якому форматі публікуються дані: CSV, JSON, PDF?

Власність та ліцензування:

- хто володіє даними?
- хто публікує дані?
- під якою ліцензією публікуються дані?
- це персональні дані? чи дані є анонімізованими?

Форма:

- як обробляються дані?
- дані існують у необробленій чи зведеній формі?

Оновлення:

- наскільки дані сучасні? як регулярно вони оновлюються?

Підтримка:

- як документується набір даних?

Для цього кожен розпорядник інформації міських рад має заповнити спеціальну анкету.

На питання в анкеті мають відповісти окремі представники кожного структурного підрозділу (департамент, управління, відділ) міських рад. Кожен структурний підрозділ (департамент, управління, відділ) міської ради має заповнити стільки анкет, скільки в їхньому розпорядженні перебуває наборів даних (1 набір даних = одна анкета).

Таблиця 1. Анкета для структурних підрозділів міських рад. Визначення наявності, формату та структури наборів даних в містах для оприлюднення у формі відкритих даних

№	Питання	Відповіді
1	Назва розпорядника інформації (департаменту, управління, відділу міської ради)	
2	Назва набору даних (наприклад, перелік об'єктів нерухомості, що здаються в оренду, або інформація щодо відстеження регулятор-них актів за 2019 рік)	
3	Ця інформація зберігається в електронному чи паперовому вигляді?	
4	Формат, в якому зараз зберігаються ці дані (doc, txt, excel, xml, csv, скан, друковані, писані, інший формат – вкажіть, який саме)	
5	Чи інформація зберігається на власних інформаційних ресурсах органу влади?	Так/Ні
6	Чи інформація зберігається у машиночитному форматі? (* про машиночитний формат – див. інструкцію).	Так / Ні
7	Як часто оновлюються дані?(наприклад, раз на місяць, раз на квартал тощо)	
8	Тип інформації (наприклад, таємно/дск/персональні дані/відкриті дані)	
9	Чи оприлюднювався набір даних до цього? Якщо відповідь "Ні" – перейдіть до питання № 14.	Так / Ні
10	Де було оприлюднено (опубліковано) дані?	
11	Початок періоду, за який було оприлюднено дані (рррр-мм-дд)	
12	Кінець періоду, за який було оприлюднено дані (рррр-мм-дд)	
13	Які формати використовувалися для оприлюднення?	
14	Чи буде інформація доступна на безкоштовній основі?	Так / Ні
15	Якщо інформація буде платною, то яка буде її вартість?	
16	ПІБ особи, відповідальної за підготовку даних	
17	Посада особи, відповідальної за підготовку даних	
18	Електронна пошта відповідальної особи	
19	Контактний телефон відповідальної особи	

Етап №2

На цьому етапі, який може відбуватися паралельно із першим, для розуміння реальної ситуації в міських радах щодо впровадження політики відкритих даних, підрозділ відповідальний за публікацію публічної інформації має заповнити окрему анкету.

Таблиця 2. Анкета для структурного підрозділу міської ради, відповідального за впровадження політики відкритих даних. Загальні питання щодо визначення політики відкритих даних в місті

№	Питання	Відповіді
1	Чи є у вас є чіткий перелік (список) даних, якими розпоряджається міська рада?	Так / Ні
2	Наскільки вичерпним і докладним на сьогодні є цей список?	
3	Чи публікується інформація у форматі відкритих даних на Єдиному державному веб-порталі відкритих даних?	Так / Ні
4	Чи функціонує окремий портал відкритих даних, розроблений за ініціативи міської ради?	Так / Ні
5	Чи існує попит на оприлюднені набори даних? Як часто їх потребують?	
6	Яким чином запити та/або зворотній зв'язок впливають на якість даних розпорядника та враховуються ним для покращення майбутніх процесів збору даних та їх публікації?	
7	Якими внутрішніми нормативними актами регулюється політика відкритих даних в вашому місті?	
8	Які технічні фактори дозволяють або заважають публікувати інформацію у вигляді відкритих даних?	
9	Чи існують в вашому структурному підрозділі системи управління даними та інформацією, які підтримують створення та публікацію відкритих даних? Які саме?	
10	Чи дозволяють системи управління даними та інформацією ефективно обмінюватися даними між структурними підрозділами?	
11	Чи місто має власні е-сервіси або інформаційно-аналітичні системи для збереження та використання наборів даних (якщо має, надайте перелік цих систем та сервісів).	
12	Які переваги, на вашу думку, надасть публікація даних у формі відкритих даних?	

Етап №3

Незалежна сторона самостійно здійснює пошук вже опублікованих наборів даних, оцінює їхню якість та структурованість. На цьому етапі до процесу бажано залучити дійсно зацікавлених осіб, наприклад громадських активістів, журналістів та/чи найактивніших користувачів. Ці особи з однієї сторони могли б конструктивно покритикувати наявні проблеми, а з іншої – саме вони прагнуть розбудувати ефективну платформу відкритих даних у місті.

Етап №4

На цьому етапі відбувається обробки анкет від кожного розпорядника, систематизується інформація, що була знайдена в мережі Інтернет, а також вивчаються питання, що пов'язані зі станом впровадження політики відкритих даних у місті. Аудит даних не лише **має констатувати** наявний стан речей, а й **запропонувати** практичні речі щодо того, як ситуацію покращити. Тому готується окремий звіт, де наводяться практичні рекомендації щодо покращення публікації органами місцевого самоврядування інформації у формі відкритих даних:

- Дається загальний огляд нормативної бази з питань відкритих даних, стану управління даними загалом.
- Пропонуються рекомендації щодо вирішення типових помилок та проблем, які пов'язані з якістю та структурою даних.
- Готуються рекомендації до кожного окремого розпорядника щодо оприлюднення наборів даних.

У звіті також може пропонуватися типова структура для ще неопублікованих наборів даних.

На цьому етапі також варто задокументувати усі досліджені набори. Іншими словами – усі заповнені анкети з етапу №1 варто перевести в одну електронну таблицю, де назви питань будуть відповідними стовпцями.

Етап №5

Наступним етапом має стати впровадження представлених рекомендацій, усунення недоліків та створення тих наборів даних, які мають бути опубліковані згідно вимог чинного законодавства. Також потрібно ухвалити необхідну нормативно-правову базу, наприклад, "**Положення про відкриті дані міської ради**".

Крім того, має бути створений спеціальний **реєстр наборів відкритих даних**, які є у розпорядженні міської ради.

Структура Реєстру наборів відкритих даних може бути наступною:

id	Унікальний ідентифікатор набору даних
title	Назва набору даних
description	Короткий опис набору даних
format	Формат набору даних (xlsx, csv, json, xml)
type	Тип даних: текстові/структуровані/графічні/відео/аудіо/архівні
creation_date	Дата створення набору даних
update_frequency	Частота оновлення набору даних
dictionary	Наявність словника (структури) набору даних
personal_data	Перелік персональної інформації, яка є в наборі
size	Розмір набору даних у мегабайтах
link	Посилання на набір даних
creator	Назва розпорядника, що є відповідальним за набір даних
contact	Способи зв'язку з відповідальною особою (телефон, e-mail)
note	Додаткова важлива інформація

Етап №6

Лише після виконання всіх попередніх етапів можлива комплексна реалізація політики відкритих даних у місті за замовчуванням. На цьому етапі публікація нових наборів даних має відбуватися **не під тиском громадськості,** а на основі **розуміння** того, що опубліковані дані принесуть користь громаді.

ПУБЛІКАЦІЯ ВІДКРИТИХ ДАНИХ

Основні принципи публікації відкритих даних

Під час публікації наборів даних важливо дотримуватися певних принципів та стандартів, якими користуються у всьому світі. Більшість із них містяться у двох документах: "**Міжнародній хартії відкритих даних**[38]", до якої приєдналася 21 держава, в тому числі Україна, та "**Восьми принципах відкритості урядових даних**[39]", які були запропоновані ще у 2007 році на зустрічі 30 активістів відкритих даних, що відбулася поблизу Сан-Франциско (про цю подію ми згадували, коли говорили про історію відкритих даних).

Проаналізувавши ці документи, загалом можна виділити шість основних постулатів публікації відкритих даних:

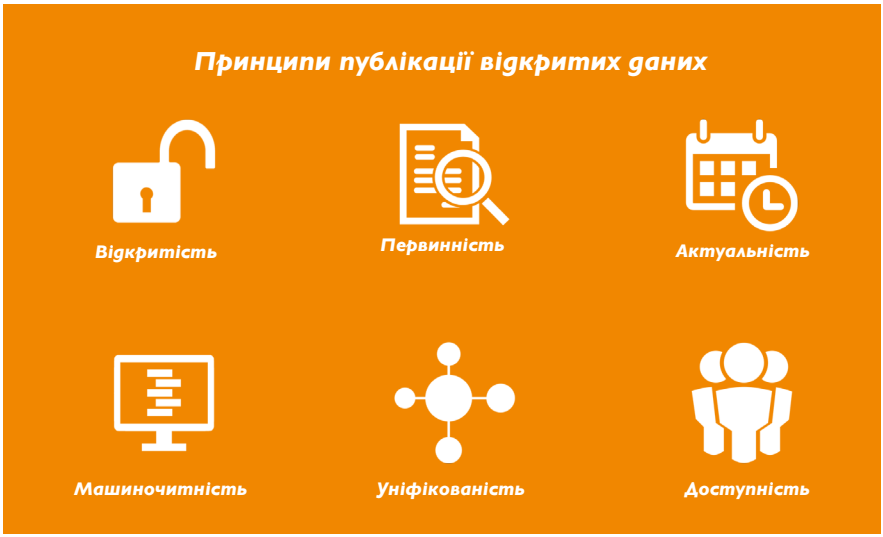
- **Відкритість за замовчуванням.** Усі публічні набори даних, що можуть бути оприлюднені – оприлюднюються. Будь-яка заборона на публікацію даних з приводу їхньої конфіденційності чи таємності має бути належно обгрунтована.
- **Первинність.** Дані, що публікуються, мають бути максимально можливо деталізованими, а не узагальненими чи агрегованими.
- **Актуальність.** Інформація публікуються максимально швидко, щоб зберегти її цінність. Крім того, наявні набори даних мають регулярно оновлюватися.
- **Машиночитність.** Набори даних мають бути у відкритих структурованих форматах даних, щоб забезпечити їхню машинну обробку.
- **Уніфікованість.** Інформація має бути максимально можливо стандартизованою. Набори даних мають використовувати однакові десяткові розділювачі, формати дати, кодування (UTF-8).
- **Доступність.** Дані розповсюджуватися безкоштовно, без жодних обмежень та контролю, на основі відкритої ліцензії та можуть використовуватись з комерційною метою.

Більше про керівні принципи ООН та країн "Великої сімки" щодо реалізації політики із відкриття державних даних українською мовою можна прочитати в ***анотації***[40] *головного консультанта відділу розвитку політичної системи Національного інституту стратегічних досліджень* ***Тетяни Джиги****.*

[38] *https://opendatacharter.net/principles*

[39] *https://public.resource.org/8_principles.html*

[40] *http://www.niss.gov.ua/content/articles/files/vidkruti_dani-6c336.pdf*



Розробка нормативної бази

Для успішної реалізації політики відкритих даних, окрім додержання національного законодавства, необхідним елементом є розробка та ухвалення відповідної нормативно-правової бази на місцевому рівні.

Створення робочої групи

Першим етапом для цього є створення робочої групи з розвитку відкритих даних. До її складу варто залучити представників IT-спеціалістів, керівників департаментів та інших зацікавлених осіб, наприклад, представників громадських організацій. Завданнями робочої групи мають бути розробка плану дій із розвитку відкритих даних, загальний моніторинг реалізації завдань, а також внесення змін до нормативно-правової бази.

План дій із розвитку відкритих даних

План дій із розвитку відкритих даних передбачає перелік завдань, часові рамки та відповідальних за їх реалізацію осіб та підрозділів. Затвердження Плану дій дасть можливість структурувати та систематизувати роботу міської ради у сфері відкритих даних.

У Плані дій варто зосередитись на таких основних завданнях:

- Нормативне забезпечення *(погодження нормативно-правової бази з нормами чинного законодавства, затвердження необхідних для налагодження процесу розпорядчих документів)*;

- Організаційно-кадрове забезпечення *(створення робочої групи, визначення/створення відповідального органу, посади чи відповідальних осіб у структурних підрозділах, державних підприємствах)*;
- Фінансове забезпечення *(передбачення коштів на розробку власного порталу відкритих даних, API до баз даних, розробка баз даних, навчання, проведення хакатонів, розробка сервісів, продуктів на основі відкритих даних)*;
- Аудит даних *(організація та проведення аудиту, впровадження рекомендацій)*;
- Контроль якості та доступності даних *(моніторинг оприлюднених наборів даних)*;
- Оприлюднення якісних даних *(визначення переліку наборів для оприлюднення, пріоритетів)*;
- Популяризація даних і співпраця з користувачами *(організація та проведення зустрічей із зацікавленими сторонами, хакатонів, опитувань)*;
- Навчання *(організація та проведення навчальних заходів, консультування)*.

Положення про відкриті дані міської ради

Головним документом у сфері відкрити даних, який має ухвалити міська влада є "Положення про відкриті дані міської ради". Варто зауважити, що документ має бути гнучким для внесення змін.

Що має містити відповідний документ?

- Основні принципи розвитку відкритих даних.
- Поняття та визначення відкритих даних.
- Аспекти нормативно-правового регулювання сфери.
- Правила публікації персональних даних, їхньої деперсоніфікації, а також обмеження через які можуть не публікуватися набори даних.
- Положення про аудит даних та правила його проведення.
- Розділ про реєстр наборів відкритих даних, зокрема, його створення, ведення та структуру.
- Формати та стандарти, які використовуються під час створення та публікації наборів даних *(формат запису дати та часу, використання десяткового розділювача)*.
- Місце публікації наборів відкритих даних *(окремий розділ на офіційному веб-сайті/створення власного Порталу відкритих даних/інтеграція з Єдиним порталом відкритих даних)*.
- Оновлюваність та використання даних.

В якості додатків до Положення бажано також долучити:

- методологію проведення аудиту даних;
- зразок Реєстру даних міської ради;
- зразок Меморандуму про співпрацю щодо відкритих даних.

До положення варто обов'язково включити принцип **відкритості за замовчуванням**. Розпорядники публічної інформації за власної ініціативи та за результатами спілкування з зацікавленими сторонами, запитами громадськості можуть публікувати додаткові (цікаві/корисні) набори даних.

У документі потрібно чітко визначити стандарти, які використовуються під час створення наборів даних. Зокрема:

- кодування UTF-8;
- міжнародний стандарт дати та часу ISO 8601 (РРРР-ММ-ДД);
- використання крапки як десяткового розділювача тощо;
- використання знаку апострофа — ' замість інших варіантів (` ´ ‘).

Також варто визначити, що структуровані дані оприлюднюються **виключно** у відкритих форматах: CSV, JSON, XML, YAML, RDF. Публікація інформації у форматах JPEG, TIFF, PDF, DOC/DOCX та інших немашиночитних форматах не допускається.

Додатково

Зразки нормативних документів для міської ради:

- Розпорядження Про заходи щодо реалізації політики відкритих даних в місті;
- Плану дій із реалізації політики відкритих даних;
- Посадових інструкцій відповідальної особи/органу за відкриті дані;
- Положення про відкриті дані міської ради;

можна знайти на Єдиному державному порталі відкритих даних України[41].

Варто розглянути можливість створення окремого органу, відповідального за відкриті дані. У такому випадку потрібно розробити Положення про орган, в якому визначити й закріпити його основні функції: впровадження, координація, контроль за відкритими даними, комунікація з користувачами, організація та проведення заходів.

Персональні та чутливі дані

Згідно закону "Про захист персональних даних" до персональних даних

[41] *https://data.gov.ua/uploads/files/2018-08-11-104314.216325Part01.pdf*

можна віднести[42] будь-які відомості, за якими ідентифікується або може бути ідентифікована фізична особа.

До таких відомостей відноситься:

- прізвище, ім'я, по батькові;
- ІНН, паспортні дані;
- адреса проживання/реєстрації;
- телефони;
- склад сім'ї;
- професія;
- освіта;
- майновий чи соціальний стан.

У цій категорії можна окремо виділити і так звані чутливі (вразливі) дані – персональні дані, обробка яких становить ***особливий ризик*** *для прав і свобод суб'єктів персональних даних.*

До таких відомостей можна віднести інформацію про:

- Расове, етнічне та національне походження;
- Політичні, релігійні та світоглядні переконання;
- Членство в політичних партіях, організаціях, професійних спілках, релігійних організаціях чи в громадських організаціях світоглядної спрямованості;
- Стан здоров'я;
- Статеве життя;
- Біометричні дані;
- Генетичні дані;
- Місцеперебування/шляхи пересування особи;
- Притягнення до адміністративної чи кримінальної відповідальності;
- Застосування щодо особи заходів в рамках досудового розслідування;
- Вчинення щодо особи тих чи інших видів насильства.

Варто відзначити, що вказані переліки **не є вичерпними**.

Захист конфіденційної інформації є важливим і складним аспектом проактивної публікації відкритих даних. Публікація таких даних може бути зроблена тільки за умови перевірки балансу: чи потенційна шкода від оприлюдненої інформації переважує суспільний інтерес в доступі до цієї інформації. У Законі

[42] *https://zakon3.rada.gov.ua/laws/show/2297-17*

України "Про доступ до публічної інформації"[43] йдеться про так званий ***"трискладовий тест"***.

Важливо усвідомити, що наявність персональної інформації в наборі даних, який готується до публікації, не означає, що цей набір не може бути оприлюдненим.

Такі набори мають пройти деперсоніфікацію.

В більшості випадків прив'язку даних до конкретної людини можна замінити ідентифікатором (не номер паспорту чи ІНН) шляхом генерації порядкового номеру, унікального ідентифікатора або GUID, що можна зробити будь-якими програмними засобами.

Наприклад, маємо набір даних щодо народжуваності у Волинській області з ПІБ матері, контактним телефоном, містом, датою народження, статтю та вагою дитини.

ПІБ	Телефон	Місто	Дата	Стать	Вага
Ігнатенко Л.Б.	332 235566	Луцьк	23.04.2018	Ж	3310
Пилипенко В.В.	332 234567	Луцьк	27.04.2018	Ж	3200
Попович А.В.	3352 13568	Ковель	01.04.2018	Ж	3325
Власенко К.В.	3352 26569	Ковель	11.04.2018	Ч	2638
Савченко К.К.	3365 35613	Ковель	23.04.2018	Ч	3423
Зорян М.О	3368 36761	Рожище	03.04.2018	Ж	3207
Коваленко О.С.	3376 27363	Маневичі	03.04.2018	Ч	3613

Деперсоналізувати ці дані, можна шляхом заміни ПІБ на унікальний ідентифікатор (не скорочення від імені). Також потрібно видалити контактний номер телефону.

ПІБ	Місто	Дата	Стать	Вага
ZS2QDUR4	Луцьк	23.04.2018	Ж	3310
VS7CKG7T	Луцьк	27.04.2018	Ж	3200
JZGQUJ4T	Ковель	05.04.2018	Ж	3325
ZZ5C8AEK	Ковель	11.04.2018	Ч	2638
SNJHBNRD	Ковель	23.04.2018	Ч	3423
NCLQJP3D	Рожище	03.04.2018	Ж	3207
W7UTE483	Маневичі	03.05.2018	Ч	3613

[43] *https://zakon.rada.gov.ua/laws/show/2939-17*

ПІБ	Місто	Дата	Стать	Вага
ZS2QDUR4	Луцьк	04.2018	Ж	3310
VS7CKG7T	Луцьк	04.2018	Ж	3200
JZGQUJ4T	Ковель	04.2018	Ж	3325
ZZ5C8AEK	Ковель	04.2018	Ч	2638
SNJHBNRD	Ковель	04.2018	Ч	3423
NCLQJP3D	Рожище	04.2018	Ж	3207
W7UTE483	Маневичі	05.2018	Ч	3613

Проте чи є дані тепер повністю анонімізованими? На жаль, ні. Бо теоретично, якщо я знаю, що моя знайома з Ковеля "Катерина Власенко" мала народити у середині квітня, я можу дізнатися стать та вагу її дитини.

Для уникнення цього, можна додатково застосувати вибіркове редагування/узагальнення даних. Наприклад, можна подати інформацію у розрізі місяців, а не конкретної дати народження.

Пам'ятайте, що для захисту конфіденційної інформації не варто відмовлятись від цілого набору даних через один проблемний елемент (або навіть декількох). Персональна інформація завжди може бути відкоригована і деперсоналізована.

Іноді деперсоналізувати дані дуже непросто. Тому, якщо Ви публікуєте набір даних з чутливою інформацією вперше – завжди можна проконсультуватися зі спеціалістами.

Можливі варіанти, коли через наявність чутливої інформації приймається рішення не публікувати набір даних взагалі. Тим не менше, ця інформація може залишатися доступною для дослідників та науковців на чітко визначених умовах.

ЗАМІСТЬ ЕПІЛОГУ

Майбутнє відкритих даних

Той, хто планує майбутнє –
планує його у свою користь.

З точки зору міжнародного досвіду, Україна лише розпочала свій шлях у світ відкритих даних. Ми, напевне, робимо багато помилок, але помилки – це завжди досвід, який допомагає нам рухатися уперед і ставати кращими.

З року в рік дедалі більше інформації оприлюднюється органами влади, а громадяни починають розуміти важливість роботи з даними та їхнього впливу на наше життя. Відкриті дані все частіше визнаються невід'ємною частиною прозорого та ефективного урядування, надання більш якісних адміністративних послуг та далекоглядного планування. Дослідники називають відкриті дані синонімом демократії. Адже це один з основних елементів для інновацій. Створюються нові продукти, послуги та сервіси. З'являються нові рішення, які допомагають робити життя громадян комфортнішим.

Тепер, наприклад, не потрібно чекати громадський транспорт незрозуміло скільки, бо завжди можна глянути, де знаходиться автобус з потрібного Вам маршруту. З іншої сторони, органи влади за допомогою відкритих даних можуть краще пояснювати свої непопулярні рішення, чим зменшують негативне ставлення до себе (звісно не завжди).

Не стоїть осторонь і бізнес. З однієї сторони компанії долучаються до створення інноваційних продуктів на основі державних даних, а з іншої – відкривають власні дані для величезної спільноти розробників. Так, в американському рейтингу Open Data 500 Companies[44] можна знайти багато відомих компаній: Amazon Web Services, Bing, Canon, сервіси Google, GitHub, Deloitte, Uber тощо.

Таким чином, від відкритих даних виграють усі: держава, громадяни, приватний сектор. І тут мова йде не лише про для багатьох абстрактні речі, як прозорість, ефективність, підзвітність, але й про цілком **матеріальні гроші**. Так, на сьогодні є декілька різних досліджень та методологій, які визначають вплив цього явища на економіку[45] в мільярди доларів. Але для того, щоб отримувати усі переваги – варто бути в тренді, відкривати дані та працювати із ними. Адже в епоху інформаційного суспільства, величезних обсягів інформації та технологій, що невпинно розвиваються, завжди виграє той, хто не лише **володіє інформацією**, а й правильно її **використовує**.

[44] *http://www.opendata500.com*

[45] *https://www.europeandataportal.eu/en/highlights/economic-benefits-open-data*

Big Data

Великі дані (англ. Big Data) — набори інформації (як структурованої, так і неструктурованої) настільки великих розмірів, що традиційні способи та підходи не можуть бути застосовані до них.

У широкому сенсі про "великі дані" говорять як про соціально-економічний феномен, що пов'язаний з появою технічних можливостей аналізувати величезні набори даних.

На сьогодні різні дослідники виділяють від 3 до 10 основних характеристик Big Data. Давайте розглянемо п'ять ключових з них, які ще називають "V's" (оскільки в англійській мові всі слова починаються з відповідної букви):

Volume (об'єм) – мабуть одна з найвідоміших характеристик Big Data. Накопичені дані настільки великі, що їх практично нереально обробляти та зберігати традиційними способами. Так, за підрахунками експертів, кожну хвилину на YouTube завантажується 300 годин відео, а щороку робиться понад 1 трлн фото.

Velocity (швидкість) – швидкість накопичення даних постійно збільшується. Наприклад, 90% всієї інформації, якою оперує людство, зібрано за останні декілька років. Також ця характеристика має на увазі швидкість обробки даних. Наприклад, Google обробляє в середньому понад 70 000 пошукових запитів щосекунди.

Variety (різноманітність) – раніше людство зосереджувалося на обробці структурованих даних. Проте насправді 80% наборів зараз є неструктурованими. Ця характеристика означає можливість одночасної обробки різних типів інформації.

Veracity (достовірність) – обсяг інформації постійно збільшується, проте чи залишаються дані достовірними? Це один з головних викликів у сфері Big Data. Хто створив дані? Хто їх редагував? Який їхній загальний контекст? Яка методологія була використана при їхньому створенні? Це лише маленький перелік питань для визначення достовірності даних, що у свою чергу допомагає краще прораховувати потенційні ризики.

Value (цінність) – мабуть, найважливіша цінність Big Data. Адже інші характеристики не мають значення, якщо Ви не можете використати ці дані. Крім того, під час збору великої кількості інформації варто одразу розуміти усі потенційні витрати та переваги кінцевого результату.

Таким чином, на сьогодні фактично необмежена велика кількість даних, яка щосекунди збільшується, дає змогу вирішувати складні глобальні проблеми: від боротьби з голодом до лікування хвороб і прогнозування надзвичайних ситуацій.

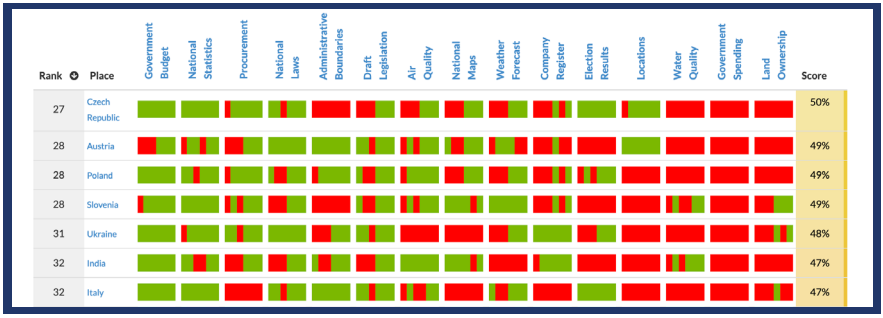
Проте, окрім беззаперечних переваг, великі дані несуть і великі ризики, які напряму пов'язані із приватністю, безпекою та недискримінацією.

Рейтинги у сфері відкритих даних

Global Open Data Index[46]

Дослідження **Open Knowledge Foundation** по вивченню відкритості державних даних у країнах світу з точки зору громадськості. Воно оцінює наявність та якість ключових наборів даних у категоріях, важливих для забезпечення прозорості діяльності та підзвітності влади.

За результатами останнього дослідження Україна знаходиться на 31 місці у рейтингу і щороку покращує свої позиції (у 2015 році було 54 місце).



Open Data Barometer[47]

Дослідження **World Wide Web Foundation**, яке оцінює наявність і якість ключових наборів даних у 15 категоріях, а також прогрес кожної країни за трьома параметрами:

- Готовність уряду, суспільства та бізнесу до відкриття даних.
- Імплементація законів про відкриті дані.
- Політичний, економічний та соціальний ефект від відкриття урядових даних.

Зараз Україна, згідно методології, набирає 47 балів зі 100, але демонструє позитивну динаміку.

[46] *https://index.okfn.org*

[47] *https://opendatabarometer.org*

Корисні посилання

- ***charted.co*** – швидка візуалізація CSV/Google Spreadsheet
- ***convertcsv.com*** – онлайн-конвертор форматів даних
- ***ukraine.apps4cities.org***– портал проекту "Дані міст"
- ***data.gov.ua/pages/infohub*** – інфохаб Єдиного державного порталу відкритих даних
- ***data.rada.gov.ua*** – відкриті дані Верховної Ради
- ***datacamp.com*** – онлайн курси по аналізу даних
- ***datajournalism.tools*** – підбір інструментів для аналізу даних
- ***fastcharts.io*** – швидка візуалізація CSV/TSV від Financial Times
- ***infogram.com*** – онлайн-інстурмент для створення інфографіки
- ***internetlivestats.com*** – онлайн статистика Інтернету
- ***mindlab.media*** – перший український журнал про відкриті дані
- ***okfn.org*** – портал організації Open Knowledge, що просуває цінності відкритих даних
- ***opendatacharter.net*** – Міжнародна хартія відкритих даних
- ***opendatahandbook.org*** – путівник по відкритих даних
- ***socialdata.org.ua/manual*** – відкритий посібник з відкритих даних УЦСД
- ***textura.in.ua*** – блог про візуалізацію даних
- ***texty.org.ua*** – аналітика та журналістика даних
- ***thedata.media*** – аналітика та візуалізація даних
- ***theodi.org*** – портал Інституту відкритих даних
- ***tidyverse.org*** – найнеобхідніші R-бібліотеки для роботи з даними

Автори:

Андрій Савчук

Руслана Величко

Дизайн, верстка — Іван Юрчик