ПІДГОТОВКА ДАНИХ ДО ПУБЛІКАЦІЇ

Публікація даних потребує ретельного планування.
Цей розділ має на меті дати розпорядникам базовий алгоритм підготовки наборів даних до оприлюднення та налагодження процесу публікації й оновлення даних на Єдиному державному порталі відкритих даних.

Перш ніж публікувати дані, потрібно провести їх ревізію. Якщо ви маєте оприлюднити набори даних, визначені Постановою Кабінету Міністрів України №835 (зі змінами), візьміть цей перелік наборів даних і складіть табличку, у якій зазначте такі пункти:

- Назва набору даних;
- Розпорядник набору даних;
- Відповідальна особа за ведення набору даних;
- Формат ведення і збереження набору даних (наприклад, автоматизована інформаційна система, файл у форматі XLS(X), один або кілька файлів у форматі DOC(X));
- Тип даних у наборі (наприклад, текстові чи структуровані);
- Частота оновлення даних;
- Наявність/відсутність у наборі персональних даних службової інформації або іншої інформації, що не підлягає оприлюдненню.

Якщо ви вже робили аудит наборів даних, це завдання не забере у вас багато часу. У іншому разі вона стане гарним початком для проведення повноцінного аудиту даних.

На цьому ж етапі вам важливо для себе з’ясувати низку питань.

Які дані потенційно мають найбільшу цінність для користувачів? Традиційно найбільш цінними для користувачів є структуровані дані у машиночитаних форматах. Саме вони зазвичай стають основою для аналітичних матеріалів, журналістських розслідувань, сервісів і застосунків. Якщо ви маєте такі набори даних, варто приділити їм першочергову увагу.

Чи відповідає формат типам даних, які в них містяться? Наприклад, ви маєте автоматизовану інформаційну систему, у якій збираються та зберігаються структуровані (табличні) дані. За таких умов із великою вірогідністю ви зможете без проблем публікувати структуровані дані у машиночитаних форматах . Або ви маєте реєстр у табличному вигляді, який ведеться у файлі формату DOC(X). У такому разі формат файлу не відповідає типу даних, що в ньому містяться, а отже вам доведеться трансформувати ці дані перед публікацією.

Який формат публікації набору даних є оптимальним? Постановою Кабінету Міністрів України №835 (зі змінами) передбачені різні формати для різних типів даних. Якщо у вашому наборі даних містяться структуровані (наприклад, табличні) дані, тоді їх треба оприлюднювати у форматі CSV. Якщо у наборі є структуровані та ієрархічні дані, для них підійдуть формати JSON і XML. Для текстових даних потрібно використовувати формат RTF. Якщо набір даних оновлюється щодня, то, можливо, для нього варто розробити інтерфейс прикладного програмування (API). У будь-якому разі вам потрібно визначити, який формат є оптимальним для кожного конкретного набору даних. А якщо формат ведення набору даних не збігається з оптимальним для нього форматом публікації, то потрібно ще й визначити процедуру трансформації даних.

Якою має бути частота оновлення даних? Згідно з Постановою №835 (зі змінами), розпорядники можуть самостійно визначати частоту оновлення оприлюднених даних. Частота може варіюватись від “відразу після внесення змін” до “щороку”. У кожному випадку варто зважати на те, як часто оновлюються дані в самому наборі (реєстрі, базі даних). Додатковим критерієм може бути наявність запиту на оновлення даних від потенційних користувачів. Наприклад, якщо робота певного сервісу, застосунку чи аналітичного інструменту потребує щотижневого оновлення даних, є сенс оновлювати дані кожного тижня.

Чи потрібно буде деперсоналізувати дані перед публікацією? Персональними даними, відповідно до закону “Про захист персональних даних”, є відомості чи сукупність відомостей про фізичну особу, яка ідентифікована або може бути конкретно ідентифікована. Наприклад, це може бути прізвище, ім’я, по батькові, дата й місце народження, місце проживання, паспортні дані й ідентифікаційний код тощо. Якщо у наборі даних, який ви маєте оприлюднити, містяться персональні дані, потрібно визначити процедуру деперсоналізації даних. Це питання також є важливим під час визначення періодичності, з якою набір даних буде оновлюватися.

Скільки часу потребує підготовка даних до публікації? Відповіді на попередні питання допоможуть вам краще зрозуміти обсяг робіт із підготовки наборів даних до публікації, визначити свої організаційні та технічні потреби, а також призначити відповідальних осіб і терміни виконання поставлених завдань.

ВИЗНАЧЕННЯ ФОРМАТУ ПУБЛІКАЦІЇ ДАНИХ

Одним із найбільш поширених питань, що виникають у розпорядників даних, є питання вибору файлового формату для публікації наборів даних. Цей розділ має на меті пояснити, як обрати файловий формат для найбільш поширених типів даних.

Постановою Кабінету Міністрів України №835 (зі змінами) передбачено публікацію декількох типів даних, як-от текстових, структурованих, геопросторових, графічних, відео- та аудіоданих, розроблених із використанням програми Macromedia Flash, архівів даних.

Важливо те, що для кожного типу даних Постановою встановлені відповідні файлові формати.

Тип даних	Формат
Текстові дані	TXT, RTF, ODT, DOC(X), PDF (з текстовим змістом, нескановане зображення), (X)HTML
Структуровані дані	RDF, XML, JSON, CSV, XLS(X), ODS, YAML

Тип даних	Формат
Геопросторові дані	GeoTIFF, SHP, DMF, MID/MIF, DXF, XML, GeoJSON, GPX, LOC, ARINC, AIXM
Графічні дані	GIF, TIFF, JPG (JPEG), PNG
Відеодані	MPEG, MKV, AVI, FLV, MKS, MK3D
Аудіодані	MP3, WAV, MKA
Дані, розроблені з використанням програми Macromedia Flash	SWF, FLV
Архів даних	ZIP, 7z, Gzip, Bzip2

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, варто мати на увазі, що під час створення нових наборів даних перевага має надаватись відкритим файловим форматам, тобто таким, що не залежать від платформи та доступні без обмежень, які можуть перешкодити їх повторному використанню. До відкритих форматів, зокрема, належать формати, як-от ODT, HTML, RDF, XML, JSON, CSV, ODS, YAML.

ТЕКСТОВІ ДАНІ

Якщо більшість місця у вашому наборі займає простий текст – ви маєте справу з текстовими даними. Прикладом текстових даних можуть бути нормативно-правові акти, положення, рішення чи розпорядження.

Публікувати текстові дані потрібно передусім у відкритих форматах RTF та ODT. Дозволяються також формати DOC(X) і PDF (з несканованим зображенням).

Категорично не підходять для текстових даних формати JPG, JPEG, PNG, GIF, TIFF, а також PDF зі сканованим зображенням. Публікація текстових даних у цих форматах унеможливлює їх обробку автоматизованими засобами.

Якщо ви плануєте публікувати багато типових наборів даних у текстових форматах, наприклад, рішень міської ради, є сенс додатково створити табличку у форматі CSV, у якій буде подано перелік цих рішень. Наприклад, зі зазначенням дати ухвалення, ідентифікаційного номеру, заголовку чи опису одним реченням та назви файлу, у якому міститься повний текст рішення чи посилання на нього.

Подібний формат публікації рішень практикує Державна судова адміністрація.

У такому разі для кожного рішення потрібно не створювати окремий набір даних на Єдиному державному порталі відкритих даних, а додавати їх у один із загальною назвою “Рішення…” або ж “Розпорядження…”.

СТРУКТУРОВАНІ ДАНІ

Якщо у вашому наборі даних є таблиці, значить, ви маєте справу зі структурованими даними. Прикладами структурованих даних можуть бути різноманітні реєстри, переліки, плани та звіти.

Публікувати структуровані дані потрібно насамперед у відкритих форматах CSV та ODS. Дозволяється також формат XLS(X).

Якщо ваші дані зберігаються в інформаційній системі чи базі даних, яка дає змогу вивантажувати дані у форматах XML чи JSON, варто використовувати їх для публікації набору

даних. Втім, зважайте на те, що найкраще формати XML і JSON підходять для ієрархічних даних. Якщо ваші дані не є ієрархічними за природою, для їх публікації буде достатньо формату CSV.

Для публікації структурованих (табличних) даних категорично не підходять формати DOC(X), RTF, PDF, JPG, JPEG, TIFF, PNG.

ГЕОПРОСТОРОВІ ДАНІ

Якщо набір даних містить інформацію про розташування певних об’єктів із зазначенням широти й довготи, або опис меж певних територій із використанням полігонів, ви маєте справу з геопросторовими даними.

Прикладом геопросторових даних можуть бути генеральні плани населених пунктів, схеми планування територій і плани зонування територій, межі виборчих округів та дільниць, відомості з Держгеокадастру, маршрути й дані про місцезнаходження громадського транспорту і т. д.

Геопросторові дані мають публікуватися передусім у відкритих форматах GeoJSON, SHP, GPX, GeoTIFF. Також можна використовувати формати DMF, MID/MIF та інші.

Для публікації геопросторових даних за поодинокими винятками не підходять формати PNG, JPG чи JPEG, призначені для графічних даних.

Деякі набори даних, наприклад, “Інформація про рекламні засоби”, “Дані про розміщення громадських вбиралень” або “Дані про місце розміщення зупинок”, можуть бути збагачені геоданими, а саме географічними координатами, що позначають точне розташування об’єкта. Однак такі набори даних цілком можна публікувати у звичайних табличних форматах CSV чи ODS.

ГРАФІЧНІ ДАНІ

Якщо набір даних є зображенням, ви маєте справу зі графічними даними. Прикладом таких даних можуть бути схеми планування територій і генеральні плани міст, але лише за умови, що вони відсутні у форматі геопросторових даних.

Якщо ви змушені публікувати геопросторові дані у файлових форматах, призначених для графічних даних (скажімо, за браком доступу до первинних геоданих), варто зазначити в метаданих географічні координати чотирьох крайніх точок на зображенні, щоб уможливити його прив'язку до карти.

Графічні дані мають публікуватись у відкритих форматах PNG, JPG чи JPEG.

АРХІВИ

Якщо ваш набір даних міститься у файлі великого розміру, або ви публікуєте багато типових файлів, що є частиною одного набору даних, є сенс використовувати для публікації архіви даних. Вони допомагають зменшити розмір набору даних і завантажити велику кількість типових файлів за один раз.

Для публікації архівів даних насамперед треба використовувати відкриті формати ZIP та 7z. Не варто використовувати для публікації архівів даних формат RAR, який є пропрієтарним.

Якщо ви використовуватимете архів даних, переконайтесь, що ви надаєте користувачам усю необхідну інформацію про вміст архіву (формат файлів, опис структури даних) у паспорті набору даних.

ІНТЕРФЕЙС ПРИКЛАДНОГО ПРОГРАМУВАННЯ (API)

Якщо ваш набір даних містить великий обсяг інформації й часто оновлюється, доступ до нього варто запровадити за допомогою інтерфейсу прикладного програмування.

Інтерфейс прикладного програмування містить набір функцій, що дає змогу отримати доступ до набору даних або певних його частин, створювати запити з різними параметрами. Він має забезпечувати можливість автоматизованого доступу до набору даних у цілодобовому режимі без вихідних.

Прикладами надання доступу до даних за допомогою інтерфейсу прикладного програмування можуть бути Єдиний державний реєстр декларацій, Єдиний веб-портал використання публічних коштів, а також набори даних Національного банку України.

У разі запровадження інтерфейсу прикладного програмування для доступу до набору даних, треба також надати детальні інструкції щодо його використання, приклади запитів зі всіма можливими параметрами та приклади відповідей на ці запити.

СТРУКТУРУВАННЯ Й ОЧИЩЕННЯ ДАНИХ

Найбільш цінними для користувачів є саме структуровані та машиночитані дані. Однак із цим типом даних традиційно виникає найбільше проблем у розпорядників даних. Цей розділ має на меті пояснити вимоги до оформлення та публікації структурованих даних.

Відповідно до Постанови №835 (зі змінами), машиночитаним вважається формат даних, який дає змогу інформаційним системам ідентифікувати, розпізнавати, перетворювати й отримувати конкретні дані без участі людини.

Інформаційні системи й системи керування базами даних дають змогу вивантажувати набори даних у структурованих і машиночитаних форматах CSV, JSON чи XML. Проблеми виникають у тих випадках, коли набори даних ведуться у простих файлах на зразок XLS(X) чи DOC(X). Зазвичай подібні набори даних необхідно спеціально готувати до публікації, тобто приводити їх до правильної табличної структури.

ПРАВИЛА СТРУКТУРУВАННЯ ТАБЛИЧНИХ ДАНИХ

Ознаки правильної табличної структури:

- Усі змінні записані у стовпчиках;
- Усі спостереження записані у рядках;
- У таблиці немає об’єднаних комірок, один запис займає лише одну комірку;
- У таблиці немає об’єднаних записів, одна комірка містить лише один запис.

Наприклад, таблиця 1 має правильну структуру, оскільки в ній змінні (region, year, population) записані у стовпчиках, а спостереження (Київ, 2015, 2888.0 тощо) записані в рядках. Кожне спостереження займає лише одну комірку, а одна комірка містить лише одне спостереження.

Таблиця 1.

region	year	population
Київ	2015	2888.0
Київ	2016	2906.6
Київ	2017	2925.8

Таблиця 2 має неправильну структуру, оскільки в ній значення змінної (year) записані як назви змінних.

Таблиця 2.

region	2015	2016	2017
Київ	2888.0	2906.6	2925.8

Таблиця 3 теж має неправильну структуру, оскільки в ній один запис – “Київ” – займає аж три комірки.

Таблиця 3.

region	year	population
Київ	2015	2888.0
	2016	2906.6
	2017	2925.8

Аналогічно, неправильним є спосіб запису даних у таблиці 4, оскільки в ньому один запис (назва змінної “population”) займає дві комірки, а в значеннях цієї змінної змішано текст числа.

Таблиця 4.

region	year	population	
		available	permanent
Київ	2015	2888.0	2846.7
Київ	2016	2906.6	2865.3
Київ	2017	2925.8	2884.5

ЧИСТОТА ДАНИХ

Однак самої лише правильної структури недостатньо для коректної обробки даних автоматичними засобами. Дані мають бути чистими.

Чистими вважаються дані, у яких:

- Немає помилок чи одруківок, зайвих символів або пропусків;
- Для запису назв, дат, чисел вживається уніфікований формат;
- Для позначення відсутніх записів використовується NA, а не “0”, “–” чи інші аналоги;
- У межах однієї змінної вживається лише один тип даних, тобто в одному стовпчику не змішуються, наприклад, числові й текстові значення.

У таблиці 5 наведено найпростіший зразок нечистих даних на прикладі вже знайомого нам набору. У змінній “region” присутні три різних способи запису назви міста Київ. У змінній “year” – два різних варіанти запису року та змішування текстових і числових значень. У змінній “population” – три різних варіанти запису числа, один з яких змішує текстові та числові значення.

Таблиця 5.

region	year	population
Київ	2015	2888.0
Києв	2016	2906.6
Кийів	‘17	2,9258 млн

КОРИСНІ ІНСТРУМЕНТИ ПЕРЕВІРКИ ДАНИХ ПЕРЕД ПУБЛІКАЦІЄЮ

Для виявлення проблем у даних ви можете використовувати безкоштовну програму Dataproofer. Вона застосовує до кожного набору даних 12 різних тестів, зокрема перевіряє наявність пропущених даних і дублікатів, підозрілі символи чи значення тощо.

Для очищення даних перед публікацією ви можете скористатись безкоштовною програмою OpenRefine. Вона дає змогу швидко виявляти різні варіанти написання назв і приводити їх до однієї форми, видаляти дублікати та зайві символи, змінювати структуру даних, заповнювати порожні комірки тощо.

ПАМ'ЯТКА ДЛЯ САМОПЕРЕВІРКИ

Погано структуровані та "брудні" дані не є машиночитаними. Їх неможливо обробляти автоматизованими засобами без участі людини. Добре структуровані й чисті дані можна швидко аналізувати, візуалізувати, використовувати в сервісах чи дослідженнях. Тож перед публікацією табличних даних переконайтеся, що:

- У таблиці немає порожніх рядків на початку або всередині;
- У першому рядку записані назви змінних латинкою (для назв змінних бажано використовувати загальноприйняті словники на зразок schema.org). У рядках із другого й до останнього записані спостереження;
- У таблиці не використовується форматування (стиль, колір чи розмір шрифтів, колір заповнення комірок);
- У межах однієї змінної вживається один тип даних. Тобто, якщо в певному стовпчику записані числові дані, у ньому не має бути текстових даних;
- У межах однієї змінної використовується один формат запису дат, назв, чисел. Для роздільника десяткових значень у числах вживається крапка (наприклад, "10.1", а не "10,1"). Для запису дат використовується формат РРРР-ММ-ДД (наприклад, 2017-12-31 для позначення 31 грудня 2017 року);
- Для позначення пропущених чи відсутніх значень використовується "NA" (а не 0, -, -99999 чи будь-яке інше значення);
- У даних нема дублікатів;
- Файл збережено з кодуванням UTF-8 у форматі CSV. Назва файлу записана латинкою.

ЗНЕОСОБЛЕННЯ ДАНИХ

Поняття персональних даних в Україні визначається Законом України "Про захист персональних даних". Відповідно до цього закону, персональними даними можна вважати "відомості чи сукупність відомостей про фізичну особу, яка ідентифікована або може бути конкретно ідентифікована".

До таких відомостей, наприклад, може належати прізвище, ім'я, по батькові, дата і місце народження, місце проживання, номер телефону, паспортні дані та ідентифікаційний код, національність, освіта, сімейний стан, релігійні та світоглядні переконання, стан здоров'я, матеріальний стан, дані про особисті майнові та немайнові відносини цієї особи з іншими особами.

Така інформація про фізичну особу та членів її сім'ї є конфіденційною і може поширюватись тільки за їхньою згодою, крім випадків, визначених законом, і лише в інтересах національної безпеки, економічного добробуту та захисту прав людини.

Не є конфіденційною інформацією персональні дані щодо здійснення посадових або службових повноважень особою, яка займає посаду, пов'язану з виконанням функцій держави або органів місцевого самоврядування.

До інформації з обмеженим доступом не належать персональні дані, зазначені в декларації про майно, доходи, витрати та зобов'язання фінансового характеру (окрім відомостей, визначених Законом України "Про засади запобігання і протидії корупції").

Також не належить до інформації з обмеженим доступом інформація, що стосується отримання в будь-якій формі фізичною особою бюджетних коштів, державного чи комунального майна, крім випадків, передбачених статтею 6 Закону України "Про доступ до публічної інформації".

Отже, перед публікацією наборів даних, треба видалити з них персональні дані, які належать до конфіденційної інформації або до інформації з обмеженим доступом.

ОФОРМЛЕННЯ НАБОРУ ДАНИХ

Оформлення набору даних під час публікації на Єдиному державному порталі відкритих даних – назва, опис і ключові слова, що вживаються для його характеристики, – є не менш важливим, ніж підготовка самих даних, їх структурування й очищення. Якісно оформлений набір даних легко знайти. Його зміст просто зрозуміти з опису та метаданих. Цей розділ містить рекомендації щодо найважливіших елементів оформлення наборів даних під час публікації.

НАЗВА НАБОРУ ДАНИХ

Назва є критично важливою для пошуку та ідентифікації набору даних на Єдиному державному порталі відкритих даних. Коректна назва дає змогу швидко знайти й ідентифікувати набір даних. Некоректна назва ускладнює знаходження набору даних і зменшує шанси на його використання.

Якщо ви публікуєте набір даних, передбачений Постановою Кабінету Міністрів України №835 (зі змінами), використовуйте назву набору даних, зазначену в Постанові. Якщо ж ви публікуєте набір даних, який не передбачений Постановою, використовуйте його офіційну назву, зафіксовану в нормативно-правових актах, або її скорочену версію.

Не використовуйте без потреби в назві набору даних назву розпорядника, місто чи регіон його розташування, дати, номери чи будь-які інші ідентифікатори. Ці відомості безумовно важливі для ідентифікації набору даних, але їх варто зазначати в інших місцях, наприклад, у паспорті набору даних.

Ідеальна назва має бути коротка й загальна. З неї має бути відразу зрозуміло, про що йдеться в наборі даних.

Приклад некоректної назви набору даних: “Річний план закупівель на 2018 рік Українського центру оцінювання якості освіти (зі змінами від 07.02.2018 року)”.

Приклад ліпшої назви набору даних: “Річний план закупівель на 2018 рік”.

Приклад коректної назви набору даних: **“**Річні плани закупівель” – за умови, що всі річні плани закупівель конкретного розпорядника даних публікуються в одному наборі даних.

ОПИС НАБОРУ ДАНИХ

Опис потрібен для того, щоб користувачі мали можливість зрозуміти, про що йдеться в наборі даних, без необхідності його завантажувати й відкривати. Він має у стислій формі надавати інформацію про зміст набору даних.

Зокрема, в описі набору даних можна зазначити часовий період, яким він обмежується, згадати адміністративно-територіальні одиниці, яких він стосується, навести перелік найважливіших змінних, які в ньому вживаються.

Опис набору даних також може містити інформацію щодо методології збору та трансформації даних. Окрім того, в описі є сенс зазначити можливі способи застосування даних, наприклад, види аналізу, який дає змогу здійснювати набір даних.

В описі не варто наводити цитати з набору даних або ж нормативно-правових актів, пов’язаних із ним.

Приклад некоректного опису набору даних: Місто Київ – освітній, науковий, політичний, культурний, спортивний та економічний центр України, місто-лідер в Україні за кількістю народжених, часткою молоді в загальній чисельності населення, кількістю студентів і внутрішніх мігрантів, які обрали місто для життя і праці та є переважно молодими людьми. Усього в місті Києві станом на серпень 2015 року мешкало 2888,3 тис. осіб, з них близько 858 тис. осіб – молодь віком 14-35 років, що становить 30% від загальної чисельності населення міста. Зважаючи на високу частку молодого населення міста, високий відносно інших міст і регіонів держави рівень соціально-економічного розвитку, а також особливий статус Києва

та його суттєве значення для побудови процвітаючої України, у місті необхідно формувати і впроваджувати активну політику щодо інтелектуального, морального, фізичного розвитку молоді, реалізації її освітнього й творчого потенціалу, спортивних та інших досягнень.

Приклад коректного опису набору даних: Набір даних містить інформацію щодо Міської комплексної цільової програми «Молодь та спорт столиці» на 2016-2018 роки. Зокрема у наборі є паспорт програми, її бюджет за різними напрямами та підпрограмами, інформація про підпрограми підтримки молоді та розвитку спорту, основні напрями діяльності й перелік заходів у межах програми, а також результативні показники її виконання.

КЛЮЧОВІ СЛОВА

Ключові слова потрібні для того, щоб набори даних можна було об'єднати в тематичні групи, за якими їх легко знайти й ідентифікувати. Ключові слова також використовуються як елементи інтерфейсу Єдиного державного порталу відкритих даних: клік на певне ключове слово дає змогу отримати всі набори даних, в описі яких воно використовується.

Власне тому як ключові варто використовувати короткі слова, які в загальній формі описують зміст набору даних.

Не варто використовувати словосполучення чи фрази, вузькі і специфічні терміни. Не треба без потреби запроваджувати нові ключові слова для опису набору даних. Насамперед варто використовувати ключові слова вже наявні на Єдиному державному порталі відкритих даних.

Приклад некоректних ключових слів: наказ, запити, запити на отримання публічної інформації, Інструкція про порядок обчислення та внесення платежів за спеціальне використання рибних та інших водних живих ресурсів, Податкова знижка, сума витрат
Класифікація товару(ів), Основні правила інтерпретації УКТЗЕД, примітки до розділів, примітки до груп, код(и) товару(ів), товарна група, позиція, підпозиція, категорія, підкатегорія, Переміщення через митний кордон, Товари, Митна вартість, Контроль, Посадові особи,

Рекомендації, екологічний податок, побутові відходи, Право на спадщину та/або посвідчені договори дарування, методичні рекомендації щодо організації та проведення органами державної фіскальної служби зустрічних звірок, обміну податковою інформацією при здійсненні податкового контролю використання води юридичними особами, фізичними особами-підприємцями та платниками єдиного податку для задоволення виключно власних питних і санітарно-гігієнічних потреб Перелік відомостей, які містять службову інформацію Державний реєстр реєстраторів розрахункових операцій (Реєстр РРО).

Приклад коректних ключових слів: НПА, наказ, постанова, рішення, ДФС.

СТРУКТУРА НАБОРУ ДАНИХ

У Постанові Кабінету Міністрів України №835 (зі змінами) структура набору даних визначається як сукупність метаданих, що містить опис складу (елементів) набору даних, їхній формат, параметри та призначення. Якщо говорити простіше, структура набору даних – це перелік усіх змінних, що у ньому є, з розшифруванням значення та зазначенням типу даних для кожної змінної.

Структура набору даних, як і опис, потрібна для того, щоб користувачі мали можливість зрозуміти, якого роду інформація є в наборі, без потреби завантажувати його та відкривати. Водночас це свого роду словник набору даних, який дає змогу користувачам краще його інтерпретувати.

Припустимо, що ми маємо такий набір даних:

region	year	population
Київ	2015	2888.0
Київ	2016	2906.6
Київ	2017	2925.8

Структура набору даних у такому разі буде мати такий вигляд:

variable	description	type
region	регіон (область, місто)	string
year	рік, за який подається інформацій	integer
population	чисельність наявного населення станом на перше січня, тис. осіб	numeric

В описі структури набору даних також можна зазначати, які значення може набувати та чи інша змінна. У такому разі структуру можна використовувати для валідації набору даних, тобто перевірки його коректності.

Наприклад, ви можете зазначити, що значення певної змінної мають бути унікальними, а значення іншої – можуть бути порожніми. Або ж ви можете вказати, що довжина значення якоїсь із текстових змінних не може перевищувати певну кількість знаків.

Структуру набору даних варто публікувати в машиночитаних форматах CSV, JSON, XML чи RDF. Для публікації структури набору даних категорично не підходять DOC(X), PDF, PNG і аналогічні їм файлові формати.

Для створення структури набору даних ви можете скористатись онлайн-сервісом Data Package Creator, створеним міжнародною організацією Open Knowledge Foundation. Він дає змогу генерувати структуру набору даних у машиночитаному форматі JSON, на основі файлу і даними, а також зазначити для кожної змінної тип даних і опис.

Якщо ви не завантажите структуру під час додавання набору даних у машиночитаному форматі (CSV, JSON, XML) на Єдиний державний портал відкритих даних, механізми порталу запропонують вам згенерувати її вручну. Вам буде запропонована проста веб-форма, у яку можна додати перелік усіх змінних, що є в наборі даних, а також їхній опис.

ОНОВЛЕННЯ НАБОРІВ ДАНИХ

Щодня центральні органи виконавчої влади й органи місцевого самоврядування продукують величезну кількість інформації. Дані в публічних реєстрах постійно змінюються та доповнюються. Отже, потрібно оновлювати й набори даних, оприлюднені на Єдиному державному порталі відкритих даних. Цей розділ має на меті надати роз'яснення щодо порядку оновлення наборів даних.

Розпорядники можуть визначати періодичність оновлення даних самостійно.

Згідно з Постановою Кабінету Міністрів України №835 (зі змінами), розпорядники інформації можуть самостійно визначати періодичність оновлення наборів даних на Єдиному державному порталі відкритих даних. Залежно від ситуації, вона може варіюватися від "відразу після внесення змін" – у разі оприлюднення даних за допомогою інтерфейсу прикладного програмування – до "щороку".

Періодичність оновлення даних на Єдиному державному порталі відкритих даних може залежати від частоти оновлення самого набору даних у розпорядника, наявності технічних і кадрових ресурсів для забезпечення оновлення наборів даних, а також наявності конкретного запиту на оновлення даних від користувачів.

Єдиний державний портал відкритих даних має інструменти для оновлення наборів даних.

Під час завантаження нової версії файлу на Єдиний державний портал відкритих даних, стара версія автоматично буде видалена і замінена новою за умови, що обидва файли мають однакові ім'я та розширення.

Якщо ви не плануєте зберігати на Єдиному державному порталі всі попередні версії набору даних, тоді під час завантаження нової версії файлу:

- Не потрібно реєструвати новий набір даних, а достатньо просто додати нову версію файлу до вже наявного набору даних;
- Назва нового файлу має бути такою ж, як і назва старого;
- Стару версію файлу не потрібно видаляти, вона автоматично буде замінена новою.

Якщо ж ви маєте на меті надати користувачам кілька версій одного набору даних, тоді під час завантаження нової версії файлу:

- Не потрібно реєструвати новий набір даних, а достатньо просто додати нову версію файлу до вже наявного набору даних;
- Назва нового файлу має відрізнятися від назви старого;
- Стару версію файлу не потрібно видаляти.

Тож, якщо ви з певною періодичністю оновлюєте набір даних, який містить, наприклад, оперативні дані про баланс газу газосховищах, варто реєструвати один набір із загальною назвою.

Неправильно: Реєструвати окремий набір даних для кожної версії: “Баланс газу 2018-01-01”, “Баланс газу 2018-01-02” тощо.

Правильно: Реєструвати один набір даних із назвою “Баланс газу”.

Якщо ви не маєте на меті зберігати кілька версій одного файлу на Єдиному державному порталі відкритих даних, тоді:

Неправильно: Публікувати в наборі даних різні версії одного файлу з різними назвами. Наприклад, “balans_2018_01_01.csv”, “balans_2018_01_02.csv”, “balans_2018[1].csv”, “balans_2018[2].csv” і т. д.

Правильно: Публікувати всі версії одного файлу з однаковою назвою, наприклад, “balans_2018.csv”.

Якщо ж ви плануєте надавати користувачам доступ до різних версій файлу, тоді:

Неправильно: Публікувати всі версії одного файлу з однаковою назвою, наприклад, “balans_2018.csv”.

Правильно: Публікувати в наборі даних різні версії одного файлу з різними назвами відповідно до певної системи іменування версій. Наприклад, Наприклад, “balans_2018_01_01.csv”, “balans_2018_01_02.csv”.

ВИКОРИСТАННЯ НАБОРІВ ДАНИХ

Публікація інформації у форматі відкритих даних має на меті уможливити її швидке оброблення автоматизованими засобами. Маючи дані у структурованому та машиночитаному форматі, журналісти зможуть створити нові цікаві історії з картами та візуалізаціями, науковці – дослідити трансформації в політиці, економіці чи суспільстві, аналітики – розробити рекомендації щодо впровадження реформ і політик, активісти – контролювати виконання органами влади своїх обов’язків, бізнес – оптимізувати свою операційну діяльність, розробники – створити корисні сервіси та застосунки.
Інакше кажучи, відкриті дані існують передусім для використання. Без використання відкриття даних втрачає сенс.

Зрозуміло, що розпорядник даних не може нікого **змусити** використовувати свої дані. Однак він може **спонукати** та **заохотити**. Є декілька способів це зробити.

ПУБЛІКАЦІЯ ЯКІСНИХ НАБОРІВ ДАНИХ, ЯКІ Є ПРЕДМЕТОМ СУСПІЛЬНОГО ІНТЕРЕСУ

Ніщо так не стимулює використання даних, як наявність деталізованих, актуальних і якісних наборів даних, які є предметом високого суспільного інтересу та на які є запит у громадськості.

Традиційно **даними, які є предметом суспільного інтересу**, вважаються дані про бюджети всіх рівнів, публічні фінанси та державні закупівлі, законодавство, реєстр компаній і бенефіціарів, земельний кадастр і адміністративні кордони, якість повітря й води та прогноз погоди. Саме ці дані можуть стати основою аналітичного дослідження, журналістського розслідування чи сервісу.

Деталізованими наборами даних можна вважати такі, що містять неузагальнену, неагреговану інформацію. Наприклад, якщо йдеться про дані щодо якості повітря, то деталізованими можна вважати дані щоденних (або й навіть щогодинних) спостережень із кожної станції, а не дані про середній рівень забруднення повітря в місті за місяць.

Для користування також важливо, щоб дані були **актуальними** та **оновлюваними**. Навряд хтось захоче робити сервіс на основі даних за позаминулий рік. Так само навряд хтось наважиться робити сервіс на даних, у регулярності оновлення яких немає впевненості.

І нарешті дані мають бути **якісними**. Під цим можна мати на увазі як структурованість та машиночитаність даних, так і наявність словників і метаданих. Якщо користувачі не зможуть розібратись у опублікованому наборі даних через відсутність словників чи метаданих, він лежатиме невживаним на Порталі відкритих даних або буде хибно витлумаченим. Аби не допустити цього, потрібно публікувати якомога більше супровідної інформації про набір даних, яка допоможе користувачам краще зрозуміти та використовувати його.

КОМУНІКАЦІЯ ІЗ КОРИСТУВАЧАМИ

Продуктивна комунікація розпорядника інформації з користувачами (журналістами, аналітиками, розробниками, представниками бізнесу) також здатна заохотити їх до використання даних.

Комунікацію з користувачами можна умовно розділити на три етапи: до, під час та після оприлюднення даних.

До оприлюднення потрібно ідентифікувати попит користувачів на дані. Виявити, які саме набори даних та в якому вигляді мають найбільший потенціал для використання.

Щоб ідентифікувати попит на дані, можна провести онлайн-опитування, організувати зустріч або круглий стіл із зацікавленими особами чи організаціями, або ж влаштувати ідеатон – подію, під час якої потенційні користувачі та представники розпорядника даних обговорюють ідеї проектів на основі даних або проблеми, які можна вирішити за допомогою даних.

Це дає змогу зосередити зусилля розпорядника даних на найбільш цінних наборах, таких, що гарантовано будуть використані для журналістського розслідування, аналітичного дослідження чи соціально корисного сервісу. Розпорядник зможе підготувати ці набори даних до публікації у такому вигляді та форматі, у яких вони потрібні користувачам, а також визначити для них оптимальну частоту оновлення.

Під час оприлюднення варто донести інформацію про цінний набір даних до всіх зацікавлених осіб: поширити інформацію про нього на офіційній сторінці розпорядника, або/та у соціальних мережах, профільних спільнотах.
Значна частина наборів даних залишаються невикористаними з тієї простої причини, що потенційні користувачі не знають про їхнє існування. Розпорядник може і має докласти зусиль до поширення інформації про оприлюднений набір даних, щоб посприяти залученню журналістів, аналітиків та розробників до його використання.

Зрозуміло, що ця стратегія має застосовуватись лише для цікавих та якісних наборів даних. Якщо ви публікуєте десятки текстових документів на день, не варто повідомляти користувачам про кожен із них.

Після оприлюднення набору даних потрібно підтримувати зв'язок із користувачами, реагувати на їхні коментарі та побажання щодо опублікованих даних, виявляти проекти, розслідування чи дослідження на основі цих даних та публічно заохочувати їх авторів.

Деякі розпорядники публікують на своїх сайтах чи сторінках у соціальних мережах інформацію про проекти, зроблені на основі їхніх даних. Це слугує визнанням для авторів проектів і водночас популяризує дані та заохочує інших до їх дослідження.

ПРОВЕДЕННЯ ХАКАТОНІВ І КОНКУРСІВ

Для заохочення використання найбільш цінних і якісних наборів даних можна організувати хакатон чи конкурс проектів на основі відкритих даних.

Хакатон – це зазвичай дводенний захід, під час якого учасники – журналісти, аналітики, програмісти, дизайнери – працюють над вирішенням якогось завдання або розробляють проект на основі даних.

Хакатон дає змогу згуртувати навколо якоїсь проблеми (або набору даних) активних і зацікавлених людей, а також за короткий термін отримати прототип соціально корисного сервісу, аналітичного дешборду або чорновий варіант журналістського розслідування.
У сфері відкритих даних хакатони є звичною справою. В Україні щороку проводять близько 10 хакатонів. На одному з них спеціалісти з data science розробляли алгоритми ідентифікації підозрілих закупівель у системі Prozorro, а на іншому активісти та розробники створювали антикорупційні інструменти для Міністерства інфраструктури.

Хакатони добре працюють і на місцевому рівні – під час одного з хакатонів у Львові кілька груп програмістів працювали над розробкою сервісів для жителів міста.

Конкурси проектів на основі відкритих даних – не менш дієвий спосіб залучення користувачів. На відміну від хакатонів, конкурси можуть тривати значно довше ніж два дні, а також проводитись онлайн.

Як хороший приклад залучення користувачів до дослідження даних можна назвати конкурс Є-Рослідування, організований командою Єдиного порталу використання публічних коштів. Результатом цього конкурсу стала поява десятків розслідувань про використання бюджетних коштів, як на національному, так і на місцевому рівні.

Також варто відзначити масштабні конкурси проектів на основі відкритих даних EGAP Challenge та Open Data Challenge. Результатом кожного з цих конкурсів стала поява корисних сервісів на основі відкритих даних, таких як Opendatabot або ж “Суд на долоні”. Ба більше, унаслідок проведення цих конкурсів сотні людей по всій країні долучилися до роботи з відкритими даними.

Матеріал підготовлено в рамках проекту «Прозорість та підзвітність в державному управлінні та послугах» / TAPAS. Створення цього матеріалу стало можливим завдяки підтримці американського народу, наданій через Агентство США з міжнародного розвитку (USAID) та за фінансової підтримки UK aid уряду Великої Британії. Зміст матерілалу не обов’язково відображає погляди Агентства USAID або уряду США чи офіційну політику уряду Великої Британії.