TEXTY.ORG.UA

Policy Brief

Як покращити якість державних даних

Андрій Газін
Texty.org.ua, 2016 рік

Короткий зміст

Оцінити якість державних даних можна за трьома основними критеріями: машиночитаність файлового формату, структурованість та чистота даних. Більше половини опублікованих на Порталі відкритих даних наборів даних представлені в немашиночитаних форматах. Типовими проблемами наборів даних, опублікованих у машиночитаних форматах, є: погане структурування даних, відсутність уніфікації форматів дат, адрес та значень, які можуть приймати певні змінні, а також відсутність описів структур даних і некоректне кодування. Задля вирішення цих проблем слід запровадити ряд валідаторів на Національному порталі відкритих даних, які будуть покликані перевіряти набори даних під час їх завантаження розпорядниками на Портал. Разом з тим слід доповнити Постанову Кабміну про відкриті дані більш чіткими інструкціями щодо оформлення наборів даних, а також переосмислити сам підхід до збору даних у тих установах, де це здійснюється шляхом ручного внесення даних у таблиці.

Вступ

21 жовтня 2015 року Кабінет Міністрів затвердив *Постанову про набори даних, які підлягають оприлюдненню у формі відкритих даних*[1]. Поміж іншим ця Постанова затверджувала перелік наборів даних, які підлягають оприлюдненню на Національному порталі відкритих даних, а також встановлювала вимоги щодо оформлення та форматів публікації наборів даних.

На виконання Постанови відводилось 6 місяців, себто до 21 квітня 2016 року на Національному порталі відкритих даних мали бути оприлюднені близько 300 індивідуальних наборів даних від органів центральної влади (всілякі реєстри та бази даних) і кілька сотень типових наборів даних — від органів як центральної, так і місцевої влади.

У травні 2016 року ми оприлюднили дослідження публікації наборів даних органами центральної та місцевої влади. У цьому досліджені ми виявили міністерства та держагенції, котрі не виконали або лише частково виконали Постанову Кабінету Міністрів про відкриті дані. За результатами цього дослідження Кабмін та Державне агентство з питань електронного урядування розіслали доручення всім органам влади, що фігурують у Постанові про відкриті дані, з вимогою надати роз'яснення щодо (не)виконання Постанови.

Нинішнє дослідження ставить за мету спробувати оцінити якість опублікованих наборів даних, виявити основні проблеми в даних та визначити шляхи покращення якості державних даних. Адже саме від якості даних зрештою залежатиме, чи будуть користуватись ними розробники, аналітики та журналісти, а отже — і чи буде в Україні повноцінна екосистема відкритих даних.

[1] Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних (http://www.kmu.gov.ua/control/uk/cardnpd?docid=248573101).

Критерії оцінки наборів даних

За якими критеріями слід оцінювати якість даних? На нашу думку, основних критеріїв оцінки якості даних три: машиночитаність файлового формату, структурованість та чистота. Узяті разом ці критерії визначають, чи можлива автоматична обробка набору даних (тобто чи можливо його використовувати у програмних скриптах для аналізу чи розробки сервісу — без додаткової ручної обробки, очищення та реструктурування).

Критерій **машиночитаності файлового формату** визначає, опубліковані дані у машиночитаному (наприклад, json, xml, csv) чи немашиночитаному (pdf, doc, jpg тощо) форматі. Критерій **структурованості** визначає, чи є дані добре структурованими (тобто зберігаються у вигляді «ключ-значення», при цьому для табличних даних діє правило «всі змінні у стовпчиках, всі спостереження у рядках, в одній комірці одне спостереження, і одне спостереження займає лише одну комірку»). Критерій **чистоти** визначає, наявні чи відсутні у даних помилки, а також чи уніфіковані дані (тобто чи використовується, наприклад, єдиний формат запису дат, адрес, власних назв і загалом значень змінних).

Як додаткові критерії оцінки якості даних можна використовувати такі: наявність метаданих, наявність опису структури даних, кодування файлу.

Маючи ці критерії оцінки, ми можемо досліджувати набори даних, опубліковані на Національному порталі відкритих даних, та визначати основні проблеми в даних та шукати шлях вирішення цих проблем і підвищення якості даних.

Оцінка за критерієм машиночитаності файлового формату

Станом на 28 серпня 2016 року на Національному порталі відкритих даних було оприлюднено 6345 наборів даних (з урахуванням того, що в одному наборі даних може бути кілька файлів, загальне число файлів на Порталі на момент оцінки — 6466).

Фактично половина — 3214 — опублікованих файлів оприлюднена у форматах, які не є структурованими — doc і docx, pdf, jpg, rtf, png і tiff. Згідно з Постановою про відкриті дані, ці файлові формати призначені для публікації текстових та графічних даних, однак фактично у них публікують у тому числі і структуровані дані.

Формат даних	Кількість наборів даних
doc/docx	2029
xls/xlsx	1547
pdf	983
csv	602
odt	326
xml	318
jpg	121
rtf	74
txt	13
html	10
ods	7
json	6
rdf	5
png	4
tiff	3

Дані в файлових форматах **doc/docx, pdf, jpg, png, tiff** за визначенням є немашиночитаними, а отже на цьому оцінка їхньої якості фактично завершується. Застосовувати критерії структурованості та чистоти до цих наборів даних нема сенсу.

Натомість є сенс або додати до Постанови про відкриті дані більш чіткі роз'яснення щодо випадків використання файлових форматів doc/docx, pdf, jpg, png, tiff, або прибрати ці формати з Постанови і заборонити таким чином публікацію даних у них.

Дані у форматах **json та xml**, навпаки, є стовідсотково машиночитаними. Традиційно дані в цих файлових форматах представляють собою вивантаження (так звані дампи) баз даних та автоматизованих аналітичних систем, а отже вони майже завжди (за відсутності помилок під час вивантаження) без зайвих зусиль підлягають автоматичній обробці. Дампи з баз даних та аналітичних систем є зазвичай добре структурованими і майже завжди уніфікованими. Однак задля стовідсоткового забезпечення якості цих наборів даних слід додати на Портал відкритих даних валідатори файлового формату (задля перевірки структури файлу) та кодування (для перевірки і примусового перекодування набору даних в UTF-8).

Справжній інтерес для дослідження, таким чином, становлять набори даних, опубліковані у файлових форматах **csv** та **xls/xlsx**. Перший формат є машиночитаним за визначенням, другі — легко конвертуються у перший. Традиційно в цих файлових форматах зберігаються невеликі реєстри, які ведуться неавтоматизованим — тобто ручним — способом. А де ручний спосіб введення даних, там і великі шанси помилки, відсутності уніфікації та поганого способу структурування даних. Саме набори даних у файлових форматах csv або xls/xlsx є сенс перевіряти на предмет структурованості та чистоти в першу чергу.

Оцінка за критеріями структурованості та чистоти. Case-study

Критерій машиночитаності файлового формату є, по суті, формальним, тобто оцінити набір даних за цим критерієм можна не відкриваючи сам набір даних. Критерії структурованості та чистоти, навпаки, є змістовними, себто для оцінки певного набору даних за цими критеріями цей набір даних слід відкрити.

Оскільки відкрити понад шість тисяч наборів даних, опублікованих на Національному порталі відкритих даних, фізично неможливо, є сенс взяти один набір даних та детально розібрати всі наявні у ньому проблеми, припустивши, що всі ці проблеми в тому чи іншому вигляді будуть присутні в більшості опублікованих наборів даних (зрештою, саме в такий спосіб працює статистичне дослідження генеральної сукупності за вибіркою).

Опис даних

Для оцінки якості ми візьмемо набір даних Державної фіскальної служби «*Реєстр виданих, призупинених, анульованих ліцензій на право роздрібної торгівлі алкогольними напоями та тютюновими виробами*»[2]. Як показало наше дослідження виконання Постанови Кабміну про відкриті дані, Державна фіскальна служба опублікувала всі необхідні набори даних (чим можуть похвалитися далеко не всі розпорядники даних, навіть із числа прогресивних міністерств).

«Реєстри виданих, призупинених, анульованих ліцензій на право роздрібної торгівлі алкогольними напоями та тютюновими виробами» ведуться територіальними органами ДФС та викладаються на Портал відкритих даних Державною фіскальною службою в архіві (zip), який містить файли з даними у машиночитаному форматі (csv)[3]. Першу версію набору даних було оприлюднено 20 квітня 2016 року, після першого оприлюднення набір даних оновлювався двічі — 25 квітня та 29 червня відповідно (однак останнє оновлення на Порталі, по суті, представляє собою посилання на сайт ДФС). Для аналізу ми використовуємо версію від 25 квітня 2016 року.

Опис структури набору даних

На сторінці набору даних відсутній файл із описом структури набору даних (тобто щонайменше переліком усіх змінних, що містяться у наборі даних). У паспорті набору даних присутні ключові слова *«Код області, Дата дії ліцензії, назва, вид, назва, адреса»,* але це не є перелік змінних у наборі даних. Тож дізнатись, який перелік змінних міститься в наборі даних, можна лише завантаживши та відкривши його (оскільки дані заархівовані, традиційний перегляд у таблиці на сторінці набору даних не працює).

Відсутність опису структури даних є типовою проблемою для всіх розпорядників даних. Найкращим вирішенням цієї проблеми буде запровадження додаткового валідатора на Національному порталі відкритих даних, метою котрого буде перевірка наявності файлу з описом структури даних. У разі відсутності такого опису є сенс пропонувати заповнити форму, з якої згодом формувати файл із описом структури. Без опису структури набору даних відмовляти в публікації набору даних на Порталі.

---

[2] Реєстр виданих, призупинених, анульованих ліцензій на право роздрібної торгівлі алкогольними напоями та тютюновими виробами (http://data.gov.ua/passport/e35e186a-f806-48d4-a120-12a404c8c2db).

[3] Реєстри кожної регіональної ДФС в архіві містяться в окремому файлі. Але разом з тим їх не 25, як можна було би припустити (всі області плюс Київ, мінус Крим та Севастополь), а 32, оскільки деякі ДФС подають реєстр ліцензій на продаж алкоголю та тютюну в окремих файлах. У цьому випадку необхідне принципове рішення — або публікувати реєстри алкогольних та тютюнових ліцензій в окремих файлах, або завжди разом.

Кодування файлів

Усі файли в наборі даних мають кодування Windows-1251, що означає: на всіх операційних системах, окрім Windows, назви та вміст файлів відображатимуться некоректно.

Зразок назв файлів у наборі даних	Зразок вмісту файлу
Реєстр_ГУ_ДФС в Одеськ_й област_.csv *Реєстр_ГУ_ДФС у _вано-Франк_вськ_й област_.csv* *Реєстр_ГУ_ДФС у В_нницьк_й област_.csv* *Реєстр_ГУ_ДФС у Волинськ_й област_.csv* *Реєстр_ГУ_ДФС у Дн_пропетровськ_й област_.csv* *Реєстр_ГУ_ДФС у Донецьк_й област_.csv*	Đåºñòð âèääíèõ, ïðèçóïèíåíèõ òà àíóëüîâàíèõ ë³öåíç³é íà ïðàâî ðîçäð³áíî¿ òîðã³âë³ àëêîãîëüíèõ íàïîÿìè òà òþòþíîâèõ âèðîáíèê ó Îäåñüê³é îáëàñò³;;;;;;;;;

Проблема з кодуванням також є типовою для майже всіх розпорядників даних. Тимчасовим вирішенням цієї проблеми може бути зазначення кодування в паспорті набору даних, аби користувачам не доводилось вгадувати. Оптимальним вирішенням проблеми є запровадження механізму примусового перекодування файлу з набором даних в UTF-8 під час завантаження на Портал відкритих даних.

Оцінка за критерієм структурованості

Нагадаємо, критерій структурованості має на меті визначити, чи є дані добре структурованими (тобто зберігаються у вигляді «ключ-значення»). При цьому для табличних даних діє вимога «всі змінні у стовпчиках, всі спостереження у рядках, в одній комірці одне спостереження, і одне спостереження займає лише одну комірку».

За цим критерієм дані ДФС можна оцінити як погано структуровані, оскільки в них деякі рядки містять змінні замість спостережень, а також назви певних змінних займають більше, ніж одну комірку. Класичний приклад обох цих помилок в одному місці — формат запису терміну дії ліцензії:

термін дії ліцензії	
дата дії з	дата дії по

Тут назва змінної «термін дії ліцензії» займає дві комірки, а значенням цієї змінної є дві інших змінних «дата дії з» та «дата дії по». Подібний формат запису є неприйнятним, оскільки унеможливлює автоматичну обробку даних.

Окрім того, поширеною проблемою у цьому наборі даних є присутність значень, які не є ані змінними, ані спостереженнями, а лише номерами стовпчиків.

Код області	Орган ліцензу-вання	Ідентифіка-ційний код СГД	Назва СГД	Реєстра-ційний номер ліцензії	Вид господар-ської діяльності	Термін дії ліцензії		Адреса здійснення діяльності	Стан ліцензії
						Дата дії з	Дата дії по		
1	2	3	4	5	6	7	8	9	10

Присутність цих записів у наборі даних суттєво ускладнює, якщо не унеможливлює автоматичну обробку даних.

Важливо також зазначити, що файли різних територіальних ДФС подекуди мають різні структури даних:

- різний перелік змінних (подекуди навіть різна кількість змінних);
- різні назви однакових по суті змінних («АТ / Назва СГ / Назва СГД / Назва суб'єкта господарювання», «Адреса здійснення діяльності / Адрес аМТ»);
- різний порядок змінних;
- різне форматування (кількість відступів на початку файлу, наявність заголовків, наявність відступів між рядком назвами змінних та власне даними тощо).

Кожна окрема із зазначених вище проблем та всі вони разом створюють перешкоди для автоматичної обробки наборів даних і роблять їх із формально машиночитаних (через файловий формат csv) — реально немашиночитаними. Частину цих проблем можна вирішити шляхом впровадження додаткових валідаторів на Порталі відкритих даних, котрі будуть перевіряти csv та xls/xlsx файли стосовно дотримання правильної структури. Разом з тим потрібно надати розпорядникам даних чіткі інструкції щодо ведення реєстрів та оформлення наборів даних, аби елімінувати проблеми зі структуруванням даних ще на етапі їх створення.

Оцінка за критерієм чистоти

Нагадаємо, критерій чистоти визначає, наявні чи відсутні у даних помилки, а також чи уніфіковані дані (тобто чи використовується, наприклад, єдиний формат запису дат, адрес, власних назв і загалом значень змінних).

За цим критерієм дані ДФС можна визначити як брудні, оскільки в них присутні граматичні помилки («ососба», «підприємцт» і тощо), а також відсутня уніфікація форматів запис адрес, дат, категоріальних («вид господарської діяльності», «стан ліцензії») та числових змінних («реєстраційний номер ліцензії»).

Так, наприклад, дати можуть записуватись щонайменше в чотирьох різних форматах: 14.11.2015, 20150301, 31.окт.15, 20.03.15.

Змінна «вид господарської діяльності» може приймати більше двох десятків різних значень (замість трьох чи чотирьох): «1», «2», «4», «А», «АЛК», «алкоголь», «Алкоголь», «АЛКОГОЛЬ», «алкогольна», «алкогольними напоями», «Алкоголь(пиво)», «АН», «П», «Пиво», «ПИВО», «роздрібна торгівля алкогольними напоями», «роздрібна торгівля алкогольними напоями (пиво)», «роздрібна торгівля алкогольними напоями (пивом)», «роздрібна торгівля тютюновими виробами», «Т», «Табак», «ТВ», «ТЮТ», «тютюн», «Тютюн», «ТЮТЮН», «тютюновими виробами».

Змінна «стан ліцензії» може приймати понад три десятки значень (замість знову ж таки трьох чи чотирьох), а в змінній «назва суб'єкту господарської діяльності» багато помилок у назві форми господарювання: «фізична особа-підприємець», «фізична особа-підприємцт», «фізичнаособа-підприємець», «фізична особ-підприємець», «фізична особа-підриємець» і т.д.

Зрозуміло, що проблеми з уніфікацією форматів дат, адрес, значень, які можуть приймати певні змінні, а також проблеми із простими граматичними помилками виникають через сам спосіб ведення реєстрів. Всюди, де дані створюються за допомоги ручного введення, неминуче будуть присутні варіативність у форматах дат, назв і значень змінних, а також граматичні помилки. Уникнути цих проблем можна хіба що змінивши сам спосіб ведення подібних реєстрів — через часткову автоматизацію та запровадження форм із заданими наперед структурами, форматами та можливими значеннями змінних.

Висновки та рекомендації

Наше дослідження демонструє, що більшість наборів даних, опублікованих на Національному порталі відкритих даних, не проходить перевірки на предмет машиночитаності файлового формату, структурованості та чистоти даних. Лише незначну частку наборів даних, опублікованих у форматах xml та json (близько 5% від загального числа наборів даних), можна вважати безумовно придатними до автоматичної обробки.

Зважаючи на це, слід переосмислити сам підхід до наповнення Національного порталу відкритих даних, змінивши фокус із кількості наборів даних на їхню якість (та зручність користування самим Порталом).

Доповнити Постанову про відкриті дані

Держагентству з питань електронного урядування слід розробити та додати до Постанови Кабміну про відкриті дані більш чіткі інструкції щодо створення наборів даних (у частині використання файлових форматів та структурування даних), з метою зменшення наборів даних, які публікуються в немашиночитаних форматах doc/docx, rtf, pdf, jpg, png, tiff. В ідеалі — в принципі вилучити з Постанови формати графічних даних, а для використання текстових форматів надати детальні роз'яснення.

Запровадити механізми валідації даних на Порталі відкритих даних

Тому ж таки Держагенству з питань електронного урядування вдосконалити Національний портал відкритих даних в частині валідації наборів даних під час завантаження розпорядниками. Кожен набір даних слід перевіряти стосовно наявності опису структури набору даних (у разі відсутності — пропонувати заповнити форму, з якої буде генеруватись опис структури даних або відмовляти у завантаженні на Портал), кодування (із автоматичним перекодування в UTF-8), сама структура файлу (наявність усіх елементів структури для xml та json, відсутність об'єднаних комірок, зайвих відступів, пропущених значень — для csv, xls/xlsx).

Встановити базові стандарти уніфікації даних

Разом з тим Держагентству з питань електронного урядування (із залученням представників розпорядників даних, розробників та аналітиків, спеціалістів Українського агентства зі стандартизації) слід розробити мінімальні стандарти та механізми уніфікації даних, спираючись на вже наявні стандарти та класифікатори.