Яка машиночитаність потрібна для доступності використання відкритих даних

Микола Кузін

2025-02-13

Abstract

Машиночитаність — це структурованість документа на логічному рівні, що уможливлює автоматизоване зчитування його структури та змісту. У сфері відкритих даних цей термін важливо розглядати разом з поняттями “доступності використання” й “простоти обробки” наборів даних — тобто приймати до уваги користувацьку перспективу. З точки зору користувача важлива не лише публікація набору у машиночитаному (структурованому) форматі, а й дотримання стандартів публікації даних. В цій статті наводжу аргументи, чому для українських розпорядників назагал важливіше сфокусуватися на користувацькій перспективі й усунути бар’єри використання наборів, які вже публікуються — й лише після цього (за потреби) дивитися в бік стандартів W3C і Інтернету Даних.

## 1 Відкриті дані: вільний доступ і доступність використання

За визначенням Open Knowledge Foundation —розробників CKAN, на якій реалізований український Портал відкритих даних — “відкритість” даних полягає в тому, що будь-хто може мати до них вільний доступ, вільно використовувати, змінювати та ділитися ними[[1]](#footnote-20).

[Однак “вільний” не завжди значить “відкритий”](https://aims.gitbook.io/open-data-mooc/unit-1-open-data-principles/lesson-1.1-what-is-open-data#id-5.-challenges): часто потрібні додаткові кроки, щоби з інформації у вільному доступі зробити дані, доступні до використання.

Доступність використання — важлива категорія. У Постанові КМУ № 835 принцип “доступності використання” відкритих даних напряму пов’язується з машиночитаним форматом оприлюднених даних. А машиночитаність — зі структурованістю, що уможливлює обробку без участі людини (власне, машинну обробку). У тому ж документі визначено перелік структурованих форматів, що використовуються для оприлюднення наборів. Сюди відносяться .csv, .xls(x), .json, .rdf та інші.

Але візьмемо набір даних “[Єдиний звіт про кримінальні правопорушення](https://data.gov.ua/dataset/8b9b1677-2407-454a-bfa7-76eb638c0ea1)”, що оприлюднюється Генеральною Прокуратурою на Порталі відкритих даних. Цей набір даних оприлюднюється у структурованому форматі .xlsx і відтак, відповідно до визначень вище, відповідає принципу доступності використання. Втім, заглянуваши всередину опублікованих ресурсів (файлів) набору, побачимо, чому про “машиночитаність” та “доступність використання” тут можна говорити досить умовно:

* Всередині однієї вкладки можуть розміщуватися кілька таблиць, одна під одною. Розрізнити їх між собою для роботи в середовищі розробки “без участі людини” складно.
* Самі таблиці використовуються для відображення ієрархічних структур даних: для прикладу, за рядком “Усього кримінальних правопорушень” слідують рядки з різнорівневими значеннями “з них”. Екселівські таблиці — не кращий спосіб відображення даних, що мають ієрархічну структуру.
* Назви колонок йдуть у кілька рядків, частина цих клітинок злита між собою, що далі ускладнює зчитування цих таблиць у середовище розробки.
* Значення колонок сумуються — це різновид дублювання даних і такі рядки треба додатково вичищати.

На перший погляд маємо суперечність: набір даних опублікований у структурованому форматі .xlsx і тому є машиночитаним, але для прочитання в середовищі розробки (для машинної обробки) потребує суттєвої участі людини і тому не є машиночитаним.

У Хартії відкритих даних, що описує шість основоположних принципів публікації наборів у форматі відкритих даних, ця суперечність знімається за умови дотримання третього та четвертого принципів. В одному наголошується на важливості використовувати структуровані (машиночитані) формати, тоді як інший вказує на необхідність дотримуватися “поширених стандартів” публікації даних.

Перед тим як детальніше сказати про користувацьку перспективу і стандарти публікації даних, розглянемо досить спеціальний контекст використання поняття машиночитаності — а саме стандарти публікації даних W3C. Оскільки і “стандарти публікації даних W3C” і “поширені стандарти публікації даних” містять слово “стандарти”, а тексти з рекомендаціями стосовно кожного з них містяться на Порталі відкритих даних, треба чітко відділити одне від одного.

## 2 **Стандарти W3C і машиночитаність як інтегрованість набору в Інтернет Даних**

Тім Бернерс-Лі розповідав, що на створення Всесвітньої Мережі його наштовхнули постійні складнощі обміну документами між різними користувачами. Оскільки це були в основному текстові документи (те, що можна прочитати), звідти маємо “гіпер-текстові” (hypertext, HT) посилання, HTTP-протокол, HTML-розмітку для структурування веб-сторінок і ще ряд засадничих для сьогоднішньої Мережі стандартів та технологій.

Згодом він написав статтю[[2]](#footnote-25), де заявив, що поряд із гіпертекстовою мережею (мережею документів), потрібна мережа даних, оскільки даних стає все більше — і всі виграють, якщо набори даних будуть пов’язані з іншими релевантними наборами по всій Всесвітній Мережі. Він сформулював поняття Linked Data, “пов’язаних [лінками] даних”, або “залінкованих даних”. На відміну від гіпертекстової мережі, де лінки (посилання) на інші документи містяться в тезі <a> з атрибутом href у HTML-документі, об’єкт мережі даних містить посилання на інші об’єкти в RDF-документі.

Так, якщо розпорядник в Україні, скажімо, КМДА, публікує набір даних про парковки в Київі, при дотриманні стандарту RDF (зокрема, використанні словників DCAT), цей набір буде залінкований з усіма наборами у світі, в яких йдеться про парковки в містах — якщо в них дотримані ті ж стандарти. RDF використовує троїсті структури (триплети) для опису спостережень:

1. Тут буде суб'єкт, конкретне паркомісце в Києві зі своїм унікальним на весь Інтернет URI-ідентифікатором;
2. предикат, який описуватиме властивість цього суб’єкта — наприклад, “Кількість паркомісць”;
3. об'єкт — це значення предиката. Скажімо, “70” (паркомісць).

Цей гарний інтерактивний граф з Linked Open Data Cloud показує, яка виглядає Інтернет даних, завдяки організаціям, які дотримуються стандартів відкритих даних WC3:

Я навмисно тут дещо заглибився в технічні деталі, щоб показати, що робота з RDF-форматом — це не просто[[3]](#footnote-27). Для повноцінного використання стандартів відкритих даних W3C розпорядникам потрібні спеціалісти. На українському Порталі відкритих даних станом на 11 грудня 2024 року з 841,643 опублікованих ресурсів (файлів), у форматі RDF опубліковано лише 25 (двадцять п’ять). При чому усі у 2018 році і в одного розпорядника, Державної прикордонної служби.

# A tibble: 25 × 2  
 mimetype created   
 <chr> <chr>   
 1 application/rdf+xml 2018-11-01T13:28:03.538971  
 2 application/rdf+xml 2018-11-02T09:00:38.201840  
 3 application/rdf+xml 2018-11-01T11:14:31.685839  
 4 application/rdf+xml 2018-11-05T10:37:34.390529  
 5 application/rdf+xml 2018-11-05T10:21:52.515944  
 6 application/rdf+xml 2018-11-05T10:39:31.313447  
 7 application/rdf+xml 2018-11-02T11:02:25.354277  
 8 application/rdf+xml 2018-11-05T10:22:31.868074  
 9 application/rdf+xml 2018-11-05T10:41:03.930371  
10 application/rdf+xml 2018-11-05T10:23:24.421427  
# ℹ 15 more rows

Source: [Article Notebook](https://mykuzin.github.io/machine-readability-and-open-data/index.qmd.html)

## 3 Поширені стандарти публікації даних і увага до користувачів

Перше, що кинулося мені у вічі, коли почав читати обговорення[[4]](#footnote-31) спільноти відкритих даних Сполученого королівства — це уточення *user stories*, тобто сценаріїв використання наборів даних різними групами користувачів. Цей же підхід є у рекомендованих країною стандартах публікації відкритих даних. Для прикладу, є рекомендований стандарт публікації табличних даних[[5]](#footnote-33), в якому на самому початку описані користувачі таких даних:

* люди, які використовують табличний редактор для базового аналізу
* аналітики, які використовують дані в статистичних програмах або застосунках для бізнес-аналітики для проведення інтерактивного аналізу
* дата-сайєнтисти, які пишуть програмне забезпечення для аналізу даних, яке завантажує та обробляє дані, наприклад, відтворювані аналітичні пайплайни
* розробники, які обробляють дані в різному програмному забезпеченні
* люди, яким потрібно швидко шукати релевантні дані перед їх аналізом за допомогою спеціалізованих інструментів

Відповідно, знаючи профілі тих, хто користуватиметься даними, стають зрозумілішіми їхні потреби, з яких вже можна сформулювати вимоги до публікації самих табличних наборів. За такого підходу доступність й простота використання справді стоїть на першому місці.

На противагу, повертаючись до згаданого набору Генпрокуратури, єдиною групою користувачів, яку можна уявити розглядаючи набір у його поточному стані — це люди, які користуються роздрукуваними листками. Втім, сам факт публікації цього набору розпорядником є важливою віхою; питання стоїть як цей набір і решту табличних наборів зробити зручнішими для використання.

В Україні є сформульовані рекомендації щодо публікації табличних даних, що роблять простішою для користувачів роботу з ними. У пам’ятці «Підготовка даних до публікації», розробленої в рамках проекту «Прозорість та підзвітність в державному управлінні та послугах» / TAPAS, влучно зазначено, що «… найбільш цінними для користувачів є саме структуровані та машиночитані дані. Однак із цим типом даних традиційно виникає найбільше проблем у розпорядників даних»[[6]](#footnote-35). Там же наведені конкретні приклади, як виглядають правильні і неправильні таблиці — і чому так.

Крім того, на Порталі відкритих даних є лаконічні рекомендації з принципами оприлюднення табличних даних (вони потребують уточнення і розширення, щоб справді бути корисними розпорядникам):

* Таблиця — це впорядкована сукупність колонок та рядків.
* Кожен рядок таблиці містить один запис.
* Кожна колонка — значення, що змінюються від рядка до рядка.
* Назви колонок розміщуються в шапці. # *тут треба ще додати, що в одному рядку*
* На перетині рядків та колонок знаходяться комірки.
* Таблиця не може містити заголовків та об’єднаних комірок.
* Колір, шрифт, інше форматування тексту та комірок не є даними.

Таким чином, в Україні проблема простоти та доступності використання структурованих даних, як бачимо, поставлена, поточна мета — засвоєння базових рекомендацій щодо публікацій наборів якомога ширшим колом розпорядників.

## 4 Висновки

Візіонерський проєкт Інтернету даних, описаний у стандартах відкритих даних W3C, вимагає від ропорядників ресурсів та інституційної спроможності для реалізації цих проєктів. Більшість ЦОВВ і окремі ОМС в Україні можуть це робити — й отримають більше видимості у світі та інтеграції з релевантними для них стейкхолдерами.

Втім, є прагматичніша ціль, від якої виграють усі користувачі даних в Україні — привести ті набори даних, що вже публікуються, у відповідність базовим рекомендаціям (стандартам) до публікації наборів.

1. <https://opendefinition.org/od/2.0/ua/> [↑](#footnote-ref-20)
2. <https://www.w3.org/DesignIssues/LinkedData.html> [↑](#footnote-ref-25)
3. Це лагідний вступ до роботи з RDF (на мові R) <https://cran.r-project.org/web/packages/rdflib/vignettes/rdf_intro.html> [↑](#footnote-ref-27)
4. <https://github.com/co-cddo/open-standards/issues/40> [↑](#footnote-ref-31)
5. <https://www.gov.uk/government/publications/recommended-open-standards-for-government/tabular-data-standard> [↑](#footnote-ref-33)
6. <https://data.gov.ua/uploads/files/2018-08-11-104337.710875Part04.pdf> [↑](#footnote-ref-35)