Яка машиночитаність потрібна для доступності використання відкритих даних

Микола Кузін

2024-12-12

Abstract

Машиночитаність — це структурованість документа на логічному рівні, що уможливлює автоматизоване зчитування його структури та змісту. У сфері відкритих даних цей термін набуває додаткових значень залежно від контекстів використання — і ці значення важливо розрізняти. У концепції Пов’язаних даних (Linked Data), запропонованої винахідником Інтернета Тімом Бернерсом-Лі, структурований формат даних є однією з технічних вимог на шляху до повної інтеграції набору в Інтернет Даних. Тут діють стандарти відкритих даних W3C. Є користувацька перспектива, реалізована, зокрема, у проєкті Frictionless Data (“обробляти й не корчитися”) де машиночитаність додатково означає простоту обробки і використання опублікованих даних. З ції точки зору важливі інші стандарти — як, наприклад, стандард публікації табличних даних. В цій статті навожу аргументи, чому для українських розпорядників наразі важливіше сфокусуватися на користувацькій перспективі й усунути бар’єри використання наборів, які вже публікуються.

## 1 Відкриті дані: вільний доступ vs 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 (розширень), 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. Ця суперечність зникає, коли детальніше дивимося на контексти значень, в яких машиночитаність може використовуватись:

1. Машиночитаність як ступінь інтегрованості набору даних в Інтернет даних[[2]](#footnote-24). У впливому проєкті Пов’язаних даних Тіма Бернерса-Лі є формальна вимога використання структурованих форматів[[3]](#footnote-26) — але простота використання окремого набору користувачами даних не є фокусом цього проєкту.
2. Машиночитаність як простота автоматизованої обробки та відповідність використовуваних форматів потребам користувачів. За цієї перспекстиви вже важливо зважати на те, хто користуватиметься наборами (є поняття сценаріїв використання — англ., *use-cases*), як прибрати непотрібні бар’єри для зчитування та подальшого користування даними.

## 2 Машиночитаність як інтегрованість набору даних у мережу Інтернет

Винахідник мережі Інтернет 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[[4]](#footnote-29), 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[[5]](#footnote-31). 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 у 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 2018-11-05T10:21:52.515944  
 5 application/rdf+xml 2018-11-05T10:37:34.390529  
 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 Машиночитаність як відповідність потребам користувачів

Перше, що кинулося мені у вічі, коли почав читати обговорення[[6]](#footnote-35) спільноти відкритих даних Сполученого королівства — це уточення *user stories*, тобто сценаріїв використання наборів даних різними групами користувачів. Цей же підхід є у рекомендованих державою стандартах публікації відкритих даних. Для прикладу, є рекомендований стандарт публікації табличних даних[[7]](#footnote-37), 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 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 на принтері листками. Втім, сам факт публікації цього набору розпорядником є важливою віхою; питання стоїть як цей набір і решту табличних наборів зробити кращими і доступнішими для користування.

В Україні є сформульовані рекомендації щодо публікації табличних даних, що роблять простішою для користувачів роботу з ними. У пам’ятці “Підготовка даних до публікації” 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»[[8]](#footnote-39). Там же наведені конкретні приклади, як виглядають правильні і неправильні таблиці — і чому так.

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. # *це* ❤
* 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 приземленіша ціль, від якої виграють усі користувачі даних в Україні — привести ті набори даних, що вже публікуються, у відповідність базовим рекомендаціям до публікації наборів.

1. <https://opendefinition.org/od/2.0/ua/> [↑](#footnote-ref-20)
2. <https://blog.ldodds.com/2015/02/20/comparing-the-5-star-scheme-with-open-data-certificates/> [↑](#footnote-ref-24)
3. <https://www.w3.org/2011/gld/wiki/5_Star_Linked_Data> [↑](#footnote-ref-26)
4. <https://www.w3.org/DesignIssues/LinkedData.html> [↑](#footnote-ref-29)
5. Це лагідний вступ до роботи з RDF (на мові R) <https://cran.r-project.org/web/packages/rdflib/vignettes/rdf_intro.html> [↑](#footnote-ref-31)
6. <https://github.com/co-cddo/open-standards/issues/40> [↑](#footnote-ref-35)
7. <https://www.gov.uk/government/publications/recommended-open-standards-for-government/tabular-data-standard> [↑](#footnote-ref-37)
8. <https://data.gov.ua/uploads/files/2018-08-11-104337.710875Part04.pdf> [↑](#footnote-ref-39)