Яка машиночитаність потрібна для доступності використання відкритих даних

Микола Кузін

2024-11-14

Abstract

Машиночитаність — широкий термін, що може позначати як технічні вимоги до оформлення веб-ресурсів, що уможливлює їхню автоматизовану взаємодію в мережі Інтернет (для прикладу, концепція 5-Star Linked Data), так і доступність (легкість) автоматизованої обробки з боку живих користувачів. За кожним з цих значень стоять різні стандарти та практики — і ці значення слід розрізняти між собою, готуючи набори відкритих даних до публікації.

## 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 .json, .csv, .xls(x) 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 парадокс: 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. Цей парадокс зникає, коли детальніше дивимося на значення, в яких машиночитаність може використовуватись:

1. Машиночитаність як ступінь інтегрованості набору даних у мережу Інтернет[[2]](#footnote-24). Сюди відносяться вимоги та рекомендації, які дозволяють веб-ресурсам правильно і повно взаємодіяти один з одним. Публікація набору даних у структурованому форматі .xlsx — це вже приклад відповідності таким вимогам[[3]](#footnote-26). Концепція *5-Star Linked Data* Тіма Бернерса-Лі, про яку скажемо нижче, є найвідомішим прикладом рекомендацій щодо такої машиночитаності.
2. Машиночитаність як чистота даних та відповідність використовуваних форматів потребам користувачів. Поняття tidy data, критерії Open Data certificate, важливість юз-кейсів/ розуміння, хто буде користуватися наборами.

## 2 Машиночитаність як інтегрованість набору даних у мережу Інтернет

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

У 2006 році сер Бернерс-Лі написав статтю[[4]](#footnote-29), де заявив, що поряд із гіпертекстовою мережею, потрібна мережа даних, бо Всесвітня мережа, поряд з документами, наповнюється все більшим обсягом даних і зрозуміти, що це за інформація без стандартів опису складно. 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.

Приклад RDF-розмітки виглядає так:

@prefix dc: <http://purl.org/dc/elements/1.1/> .  
@prefix ex: <http://example.org/> .  
  
ex:book1  
 dc:title "Linked Data для початківців" ;  
 dc:creator ex:author1 ;  
 dc:subject "Semantic Web" .  
  
ex:author1  
 dc:name "Іван Коваленко" .

Якщо не дотримуватись, нічого не поламається, але важко знаходити.

Цей гарний інтерактивний граф з Linked Open Data Cloud

Зв’язані дані (переклад терміну Linked Data, який дещо губить контекст оскільки будь-які інші зміни неможливі без

### 2.1 а. 5-Star Linked

## 3 Машиночитаність як відповідність потребам користувачів

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)