Engineer

Метадані

дані, що описують інші дані, аби їх можна було знайти

Що це таке

Метадані — це дані, предметом яких є інші дані: назва, дата, мова, автор, формат і розмір, що супроводжують документ, аби люди й програми могли його впорядкувати, довіритися йому й вирішити, що з ним робити, ще до того, як прочитають сам вміст. У щоденній роботі метадані мають вужчі назви залежно від того, де вони живуть: поля заголовків у HTTP, властивості в офісному файлі, теги у фотографії, front matter на початку сторінки Markdown. В англомовних стандартах старшого покоління трапляється написання meta-data через дефіс.

Метадані зазвичай ділять на три види. Описові кажуть, про що йдеться: назва, автор, тема, анотація. Структурні пояснюють, як частини складаються в ціле: розділи, сторінки, порядок читання. Адміністративні фіксують, звідки річ походить і що з нею можна робити: коли її створено й змінено, ким, за якою ліцензією і в якому форматі. Найвідоміший спільний перелік описових полів — Dublin Core: його п'ятнадцять елементів 2003 року стали стандартом ISO 15836, а назва, автор, дата, мова й права звідти повторюються під іншими іменами майже в кожному пізнішому форматі.

Вебсторінка несе метадані шарами. Першою надходить відповідь HTTP. За RFC 9110 поле заголовка може змінювати чи розширювати семантику повідомлення, описувати відправника, визначати вміст або давати додатковий контекст, тож `Content-Type`, `Content-Language` і `Last-Modified` описують сторінку ще до того, як надійде її перший байт. Усередині HTML розміщені `<title>`, атрибут `lang` та елементи `<meta>`, імена яких визначає стандарт HTML, як-от `description`, `author` і `generator`. Над ними лежать словники, які читають інші сторони: властивості Open Graph для попереднього перегляду посилань і структуровані дані для пошукових систем та асистентів.

То що ж відрізняє метадані від даних? Радше читач, ніж самі байти. Каталожний запис — це метадані про книжку, але для каталогу він і є набором даних; час зйомки фотографії — дрібниця для того, хто її розглядає, і сировина для того, хто складає мапу місць, де фотографують. Питання «метадані чи структуровані дані» вужче, і відповідь на нього така: структуровані дані — це метадані, записані спільним словником, щоб програма могла прочитати поле, не вгадуючи, що воно означає.

Застосуйте поняття до рішення

Застосуйте «Що таке метадані?» до конкретного рішення. Ми можемо пов'язати поняття із системою, яку ви створюєте або купуєте. Правила й перевірки, що відсіюють погані дані ще на вході, і чітка відповідь, хто володіє якою цифрою і звідки вона взялася.

Переглянути наші послуги Зв'язатися з нами →

Перевірте припущення за терміном

Використайте пояснення та пов'язані роботи, щоб сформулювати запитання для проєкту. Ми можемо перевірити припущення разом до вибору інструмента. Один інфраструктурний план для 14 підрозділів — 14 дослідницьких підрозділів, охоплених спільним інфраструктурним планом

Один інфраструктурний план для 14 підрозділів Зв'язатися з нами →

Чому це важливо

Кожне поле — маленька обіцянка, і справжнє рішення полягає в тому, звідки кожне з них бере значення: від людини, яка його вписала, з історії версій, зі стовпця бази даних чи з того, що випадково повідомила файлова система. Програми діють на підставі цих обіцянок, не перевіряючи їх. У цьому і сенс метаданих, і їхній ризик.

Кеші та пошукові роботи розподіляють за ними свої зусилля. Браузер, що має копію сторінки, надсилає назад отриману дату `Last-Modified` і, якщо сторінка не змінилася, дістає коротку відповідь `304 Not Modified` замість усієї сторінки. RFC 9110 просить сервер надсилати цю дату тоді, коли останню зміну можна визначити обґрунтовано й послідовно, і ця умова працює в обидва боки: де послідовної дати немає, стандарт її й не вимагає. Дата зміни в карті сайту виконує ту саму роботу для пошукового робота, який вирішує, що завантажити знову.

Від метаданих залежать і люди, і допоміжні технології. Атрибут `lang` підказує екранному читачеві, за якими правилами вимовляти текст, а назву показують вкладка, закладка й результат пошуку. Права подорожують так

само: менеджер пакетів знає ліцензію залежності лише тому, що залежність сама її оголосила, тож перевірка ліцензій, побудована на цьому переліку, рівно настільки надійна, наскільки правдиві оголошення під нею.

Кожне поле треба підтримувати правдивим, тож сторінка із сорока полями має сорок нагод помилитися. Деякі поля не читає ніхто: стандарт HTML зазначає, що багато пошукових систем не зважають на ім'я `keywords`, бо його використовували для спаму в результатах пошуку. До того ж метадані протікають: місце зйомки фотографії чи історія правок документа можуть сказати більше, ніж сам вміст. Кілька полів, кожне з яких заповнюється з джерела, що не може розійтися з дійсністю, зазвичай кращі за багато.

Як це виглядає на практиці

Метадані трапляються у трьох проєктах на цьому сайті, щоразу в іншому виді файлу: у карті сайту, у наборі PDF і в графі залежностей.

Одна карта сайту стверджувала, що 172 зі 176 сторінок змінилися тієї самої миті. Дати бралися з часу зміни файлів, а експорт вмісту переписує всі файли за кожного запуску. Тепер дата проходить ланцюжок джерел: спершу явне поле, тоді історія комітів і лише наостанок файл. У ланцюжку є пастка, яку проєкт записав: назва поля мусить стояти в переліку, інакше генератор поле взагалі не читає. База даних за сторінками окремо зберігає, коли запис написано й переглянуто, і окремо — роки, коли відбувалася сама робота, а якщо дати ніхто не знає, поле лишається незаповненим: точне хибне число гірше за відсутнє. На зафіксованому етапі реалізації повторна синхронізація після місяця правок зачепила 8 файлів зі 186.

На зафіксованому етапі перевірки дев'ять згенерованих PDF-документів, перевірених за стандартом доступності пройшли всі 106 правил налаштованого профілю доступності. Робота включала оголошену мову документа й справжню назву в метаданих, а також теговану структуру, альтернативний текст і явно заданий порядок читання. У тому зафіксованому прогоні ті самі документи не пройшли рівно одне правило зі 146 в архівному профілі. Сторінка називає це результатом, близьким до відповідності, і вказує, яке правило не пройдено, адже 145 успішних перевірок не означають відповідності.

А коли ліцензії 644 крейтів Rust перевіряли за кожного збирання, заявлена межа перевірки виявилася межею метаданих. Перевірка читає оголошену ліцензію кожного крейта, оголошення бувають хибними або неповними, і про крейт, що оголошує себе неправдиво, перевірка нічого не доводить. Її двічі навмисно змусили зупинити збирання, і це показує, що вона працює; а от чесність оголошень, які вона читає, не здатна показати жодна перевірка.

Де виникають проблеми

У збоїв нижче одна спільна риса: кожна перевірка, яка їх не шукає, проходить.

Після кожного розгортання кожна сторінка повідомляє про зміну. Вебсервер, що надсилає час зміни статичного файлу як `Last-Modified`, або генератор, що так само датує карту сайту, повідомляє про мить, коли файл скопіювали, а клонування репозиторію, експорт чи збирання контейнера копіюють усі файли одразу. Кеші перевіряють усе заново, а пошукові роботи нічого не дізнаються. Беріть дати з історії версій або з поля, яке вписала людина, а невідому дату залишайте відсутньою. У Hugo джерела — це перелік, який читають по черзі, і назва самого поля мусить у ньому бути:

```
enableGitInfo = true

[frontmatter]
date = ["date", ":git", ":fileModTime"]
lastmod = ["lastmod", ":git", ":fileModTime"]
```

Назва або попередній перегляд посилання показує спотворені символи. Кодування символів — це метадані про всі інші байти, і парсер, що натрапляє на текст раніше за оголошення, мусить вгадувати. Стандарт HTML вимагає, щоб елемент з оголошенням повністю вміщувався в перші 1 024 байти, а якщо поставити його першим у `<head>`, вгадувати не доведеться. На цьому сайті інструмент URL Inspection від Bing якось прочитав назву й обидва рядки Open Graph як Latin-1, бо довге тире в назві трапилося раніше за оголошення `charset`. Заголовок HTTP теж називав UTF-8, але інструмент, що читає файл без заголовка, може покладатися лише на саму сторінку. Тепер `charset` стоїть першим елементом, а перевірка контролює цей порядок на кожній зібраній сторінці.

Той самий факт у різних місцях каже різне. Назва сторінки живе в `<title>`, в `og:title`, у її структурованих даних, а для PDF — ще й у властивостях документа; мова живе в `lang`, у `Content-Language` і в `og:locale`. Коли кожне значення вписують окремо, вони розходяться, і кожен споживач вірить своєму. Обчислюйте кожне значення один раз, в одному шаблоні чи одній функції, і нехай кожне місце, що його виводить, читає саме це значення.

Файл розповідає більше, ніж автор збирався оприлюднити. Фотографії несуть модель камери, час і часто координати; офісні документи — імена авторів, історію правок і шляхи до шаблонів; тег `generator` називає програму, яка зібрала сторінку. Під час експорту вилучайте все, що ніхто свідомо не вирішив публікувати, і ставтеся до кожного поля як до тексту, який прочитає незнайомець.

Що почитати далі

- RFC 9110, section 6.3: Header Fields — як HTTP передає метадані попереду вмісту і що може робити поле заголовка.
- RFC 9110, section 8.8.2: Last-Modified — коли сервер має надсилати дату зміни і чому дату з майбутнього замінюють.
- HTML Standard: the meta element — стандартні імена метаданих і правила оголошення кодування символів.
- DCMI Metadata Terms — п’ятнадцять елементів Dublin Core і словник, збудований навколо них.
- The Open Graph protocol — чотири властивості, обов’язкові для кожної сторінки, і необов’язкові навколо них.

Словник

Обирайте за фактичними умовами

Надайте репрезентативний набір даних, запитання до нього та поточне вузьке місце. Узгодимо вимірювання якості, пропускної здатності та відновлення.

Дивитися, скільки це коштує Зв'язатися з нами →

Зробіть термінологію практичною

Запитайте, як «Що таке метадані?» стосується вашої організації. Конкретного прикладу з поточної системи достатньо для початку розмови.

Переглянути наші послуги

welcome@engineer.company +45 50 10 61 84 Копенгаген, Данія

Візьміть цю сторінку із собою

Візитівка

Дбайливо зроблено з любов'ю — D.

CVR 45969703 · Tagensvej 105C, 3. th, 2200 København N

Авторське право © 2025 – дотепер · Engineer ApS · Усі права захищено.

Надруковано з https://engineer.company/uk/glossary/metadata/