Engineer

Metadaten

Daten, die andere Daten beschreiben und auffindbar machen

Was es ist

Metadaten sind Daten über andere Daten: Titel, Datum, Sprache, Autor, Format und Größe, die ein Dokument begleiten. So können Menschen und Programme es sortieren, ihm vertrauen und darauf reagieren, bevor sie es lesen. Ältere Standards schreiben meta-data mit Bindestrich. Im Alltag heißen sie enger gefasst Headerfelder in HTTP, Eigenschaften einer Office-Datei, Tags eines Fotos oder Front Matter am Anfang einer Markdown-Seite.

Eine verbreitete Einteilung unterscheidet drei Arten. Beschreibende Metadaten sagen, worum es geht: Titel, Ersteller, Thema, Zusammenfassung. Strukturelle Metadaten sagen, wie die Teile zusammengehören: Kapitel, Seiten, Lesereihenfolge. Administrative Metadaten sagen, woher etwas stammt und was damit geschehen darf: wann und von wem es erstellt und geändert wurde, unter welcher Lizenz und in welchem Format. Der Dublin Core mit fünfzehn Elementen, der 2003 zum Standard ISO 15836 wurde, ist die bekannteste gemeinsame Liste beschreibender Felder. Seine Angaben zu Titel, Ersteller, Datum, Sprache und Rechten erscheinen seither unter anderen Namen in fast jedem Format wieder.

Eine Webseite trägt Metadaten in Schichten. Zuerst kommt die HTTP-Antwort. RFC 9110 sagt, ein Headerfeld könne „die Nachrichtensemantik ändern oder erweitern, den Absender beschreiben, den Inhalt festlegen oder zusätzlichen Kontext liefern“. `Content-Type`, `Content-Language` und `Last-Modified` beschreiben die Seite, bevor ein Byte davon ankommt. Im HTML stehen `<title>`, das Attribut `lang` und `<meta>`-Elemente mit vom HTML-Standard definierten Namen wie `description`, `author` und `generator`. Darüber kommen Vokabulare, die andere Beteiligte lesen: Open-Graph-Eigenschaften für Linkvorschauen und strukturierte Daten für Suchmaschinen und Assistenten.

Was trennt also Metadaten von Daten? Eher der Leser als die Bytes. Ein Katalogeintrag besteht aus Metadaten über ein Buch und bildet für den Katalog zugleich den vollständigen Datensatz. Der Aufnahmezeitpunkt eines Fotos ist für den Betrachter ein Detail und für jemanden, der Aufnahmeorte kartiert, das Rohmaterial. Enger gefasst ist die Frage nach dem Unterschied zwischen Metadaten und strukturierten Daten. Die Antwort ist, dass strukturierte Daten Metadaten in einem gemeinsamen Vokabular sind. Ein Programm kann dadurch ein Feld lesen, ohne seine Bedeutung erraten zu müssen.

Das Konzept für eine Entscheidung nutzen

Nutzen Sie „Was sind Metadaten?“ für eine konkrete Entscheidung. Wir können helfen, das Konzept mit dem System in Verbindung zu bringen, das Sie entwickeln oder einkaufen. Regeln und Prüfungen fangen fehlerhafte Daten beim Eingang ab und klären, wer für welche Zahl verantwortlich ist und woher sie stammt.

Unsere Leistungen entdecken Kontakt aufnehmen →

Die Annahmen hinter dem Begriff prüfen

Nutzen Sie die Erklärung und die verlinkten Arbeiten, um Fragen für Ihr Projekt zu formulieren. Wir können die Annahmen gemeinsam mit Ihnen prüfen, bevor ein Werkzeug ausgewählt wird. Ein Infrastrukturplan für 14 Abteilungen — 14 Forschungsabteilungen mit einem gemeinsamen Infrastrukturplan

Ein Infrastrukturplan für 14 Abteilungen Kontakt aufnehmen →

Warum es wichtig ist

Jedes Feld ist ein kleines Versprechen. Entscheidend ist, woher es seinen Wert bekommt: von einer tippenden Person, aus der Versionsgeschichte, einer Datenbankspalte oder dem, was das Dateisystem gerade meldet. Programme handeln nach diesen Versprechen, ohne sie zu prüfen. Das ist der Zweck von Metadaten und ihr Risiko.

Caches und Crawler richten ihren Aufwand danach aus. Ein Browser mit einer gespeicherten Seitenkopie sendet das erhaltene Datum aus `Last-Modified` zurück. Hat sich nichts geändert, erhält er eine kurze Antwort `304 Not Modified` statt erneut der Seite. RFC 9110 fordert einen Server auf, dieses Datum für jede Repräsentation zu senden, deren letzte Änderung „vernünftig und konsistent bestimmt werden kann“. Die Bedingung gilt in beide Richtungen: Gibt es kein konsistentes Datum, fordert der Standard keines. Das Änderungsdatum einer Sitemap erfüllt denselben Zweck für einen Crawler, der entscheidet, was er erneut abruft.

Auch Menschen und assistive Technik hängen davon ab. Das Attribut `lang` sagt einem Screenreader, welche Ausspracheregeln er anwenden soll. Der Titel erscheint im Tab, im Lesezeichen und im Suchergebnis. Rechte werden genauso

weitergegeben: Ein Paketmanager kennt die Lizenz einer Abhängigkeit nur, weil sie diese angegeben hat. Eine auf dieser Liste beruhende Lizenzprüfung ist deshalb genauso gut wie die zugrunde liegenden Angaben.

Jedes Feld muss wahr bleiben. Eine Seite mit vierzig Feldern bietet deshalb vierzig Gelegenheiten für Fehler. Manche Felder liest niemand: Der HTML-Standard merkt an, dass viele Suchmaschinen den Namen `keywords` nicht berücksichtigen, weil er zum Spammen von Ergebnissen benutzt wurde. Metadaten geben außerdem Informationen preis. Der Standort eines Fotos oder die Überarbeitungsgeschichte eines Dokuments kann mehr verraten als sein Inhalt. Wenige Felder, deren Werte jeweils aus einer gegen Abweichungen abgesicherten Quelle stammen, sind gewöhnlich besser als viele.

Wie es in der Praxis aussieht

Metadaten treten in drei Projekten dieser Website auf, jedes Mal in einer anderen Dateiart: einer Sitemap, einer Gruppe von PDFs und einem Abhängigkeitsgraphen.

Die Sitemap einer Website bezog ihre letzten Änderungszeitpunkte aus den falschen Metadaten, sodass 172 ihrer 176 Seiten am Ende denselben Zeitpunkt teilten. Es waren Änderungszeiten von Dateien, also Metadaten über die Bytes statt über den Inhalt. Jeder Lauf des Inhaltsexports schrieb sie sämtlich neu. Die Korrektur liest eine Rückfallkette: zuerst ein ausdrücklich gesetztes Feld, danach die Commit-Geschichte, dann die Datei. Sie hält auch die darin liegende Falle fest: Der wörtliche Feldname muss in der Liste stehen, sonst liest der Generator das Feld überhaupt nicht. Die Datensätze hinter den Seiten halten jetzt den Tag, an dem ein Eintrag geschrieben und überarbeitet wurde, von den Jahren der eigentlichen Arbeit getrennt. Ein unbekanntes Datum bleibt leer, nach der Regel „Eine präzise, aber falsche Zahl ist schlimmer als eine fehlende“. Im dokumentierten Implementierungsstand betraf eine erneute Synchronisierung nach einem Monat Bearbeitungen 8 von 186 Dateien.

Im dokumentierten Validierungsstadium bestanden neun generierte PDF-Dokumente, die am Barrierefreiheitsstandard gemessen wurden, alle 106 konfigurierten Regeln des Barrierefreiheitsprofils. Zur Arbeit gehörten eine angegebene Dokumentsprache und ein echter Titel in den Metadaten sowie getaggte Struktur, Alternativtext und eine ausdrückliche Lesereihenfolge. In diesem dokumentierten Lauf scheiterten dieselben Dokumente im Archivprofil

an genau einer von 146 Regeln. Die Seite bezeichnet das unter Nennung der fehlgeschlagenen Regel als knapp verfehlt, denn 145 bestandene Regeln bedeuten keine Konformität.

Und als die Lizenzen von 644 Rust-Crates bei jedem Build geprüft wurden, war die erklärte Grenze der Prüfung eine Grenze von Metadaten. Sie liest die von jedem Crate angegebene Lizenz. Angaben können falsch oder unvollständig sein, und die Prüfung beweist nichts über ein Crate mit einer falschen Angabe. Die Prüfung wurde zweimal absichtlich zum Scheitern gebracht. Das zeigt, dass sie funktioniert. Nichts, was eine Prüfung tut, kann zeigen, dass die gelesenen Angaben ehrlich sind.

Wo es schiefgeht

Die folgenden Fehler haben eine Gemeinsamkeit: Jede Prüfung, die nicht nach ihnen sucht, meldet Erfolg.

Jede Seite meldet nach jeder Auslieferung eine Änderung. Ein Webserver, der die Änderungszeit einer statischen Datei als `Last-Modified` sendet, oder ein Generator, der eine Sitemap so datiert, meldet den Kopierzeitpunkt der Datei. Ein Checkout, Export oder Containerbuild kopiert sie alle. Caches prüfen alles erneut auf Gültigkeit, und Crawler erfahren nichts. Beziehen Sie Datumsangaben aus der Versionsgeschichte oder einem von einem Menschen geschriebenen Feld. Lassen Sie ein unbekanntes Datum fehlen. In Hugo werden die Quellen als Liste der Reihe nach gelesen, und der eigene Name des Felds muss darin stehen:

```
enableGitInfo = true

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

Ein Titel oder eine Linkvorschau zeigt kaputte Zeichen. Die Angabe zur Zeichencodierung ist eine Metainformation über jedes andere Byte. Ein Parser, der vor der Deklaration auf Text trifft, muss raten. Der HTML-Standard verlangt, dass das deklarierende Element vollständig in die ersten 1.024 Bytes passt. Steht es zuerst in `<head>`, muss nichts erraten werden. Auf dieser Website las Bings URL Inspection einmal den Titel und beide Open-Graph-Zeichenfolgen als Latin-1, weil ein Geviertstrich im Titel vor dem Charset ankam. Auch der HTTP-Header

nannte UTF-8. Ein Werkzeug, das die Bytes ohne Header liest, hat aber nur die Seite als Grundlage. Das Charset ist jetzt das erste Element, und eine Prüfung bestätigt diese Reihenfolge auf jeder gebauten Seite.

Dieselbe Tatsache sagt an verschiedenen Stellen etwas anderes. Der Titel einer Seite steht in `<title>`, in `og:title`, in ihren strukturierten Daten und bei einem PDF in den Dokumenteigenschaften. Ihre Sprache steht in `lang`, in `Content-Language` und in `og:locale`. Getrennt eingegeben driften die Werte auseinander, und jeder Verbraucher glaubt einem anderen. Berechnen Sie jeden Wert einmal in einem Template oder einer Funktion und lassen Sie jede ausgebende Stelle diesen Wert lesen.

Eine Datei sagt mehr, als ihr Autor veröffentlichen wollte. Fotos enthalten Angaben zur Kamera, zum Zeitpunkt und oft zum Standort. Office-Dokumente enthalten Autorennamen, Bearbeitungsgeschichte und Template-Pfade. Ein Tag `generator` nennt die Software, die eine Seite gebaut hat. Entfernen Sie beim Export alles, dessen Veröffentlichung niemand gewählt hat, und behandeln Sie jedes Feld als Text, den ein Fremder lesen wird.

Weiterführende Lektüre

- RFC 9110, Abschnitt 6.3: Headerfelder — wie HTTP Metadaten vor dem Inhalt überträgt und was ein Headerfeld tun kann.
- RFC 9110, Abschnitt 8.8.2: Last-Modified — wann ein Server ein Änderungsdatum senden sollte und warum ein zukünftiges ersetzt wird.
- HTML-Standard: das meta-Element — die standardisierten Metadatennamen und Regeln für die Angabe einer Zeichencodierung.
- DCMI-Metadatenbegriffe — die fünfzehn Dublin-Core-Elemente und das darum aufgebaute Vokabular.
- Das Open-Graph-Protokoll — die vier auf jeder Seite erforderlichen Eigenschaften und die optionalen daneben.

Glossar

Für Ihre tatsächlichen Rahmenbedingungen auswählen

Bringen Sie einen repräsentativen Datensatz, die Fragen, die er beantworten muss, und den aktuellen Engpass mit. Wir können vereinbaren, wie Qualität, Durchsatz und Wiederherstellung gemessen werden.

Preise ansehen Kontakt aufnehmen →

Die Begriffe praktisch nutzbar machen

Fragen Sie, wie „Was sind Metadaten?“ auf Ihre Organisation zutrifft. Ein konkretes Beispiel aus Ihrem bestehenden System reicht aus, um das Gespräch zu beginnen.

Unsere Leistungen entdecken

welcome@engineer.company +45 50 10 61 84 Kopenhagen, Dänemark

Nehmen Sie diese Seite mit

Kontaktkarte

Mit Sorgfalt und Liebe gestaltet von D.

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

Copyright © 2025 – heute · Engineer ApS · Alle Rechte vorbehalten.

Gedruckt von https://engineer.company/de/glossary/metadata/