Engineer

Metadata

de data der beskriver andre data, så de kan findes

Hvad det er

Metadata er data, hvis emne er andre data: titlen, datoen, sproget, forfatteren, formatet og størrelsen, der følger med et dokument, så mennesker og programmer kan sortere det, stole på det og handle på det, før nogen har læst indholdet. I det daglige har metadata smallere navne alt efter, hvor de bor: headerfelter i HTTP, egenskaber i et kontordokument, tags i et foto og front matter øverst i en Markdown-side. Ældre engelske standarder staver ordet meta-data med bindestreg.

Man skelner gerne mellem tre slags. Beskrivende metadata fortæller, hvad noget handler om: titel, ophav, emne og resumé. Strukturelle metadata fortæller, hvordan delene hænger sammen: kapitler, sider og læserækkefølge. Administrative metadata fortæller, hvor noget kommer fra, og hvad man må gøre med det: hvornår det blev oprettet og ændret, af hvem, under hvilken licens og i hvilket format. Den bedst kendte fælles liste over beskrivende felter er Dublin Core, hvis femten elementer blev standarden ISO 15836 i 2003, og titel, ophav, dato, sprog og rettigheder derfra går igen under andre navne i næsten alle senere formater.

En webside bærer metadata i lag. Først kommer HTTP-svaret. Ifølge RFC 9110 kan et headerfelt ændre eller udvide beskedens betydning, beskrive afsenderen, definere indholdet eller give ekstra kontekst, og `Content-Type`, `Content-Language` og `Last-Modified` fortæller om siden, før dens første byte er nået frem. I selve HTML'en ligger `<title>`, attributten `lang` og de `<meta>`-elementer, hvis navne HTML-standarden fastlægger, for eksempel `description`, `author` og `generator`. Oven på dem kommer ordforråd, som andre parter læser: Open Graph-egenskaber til linkforhåndsvisninger og strukturerede data til søgemaskiner og assistenter.

Hvad adskiller så metadata fra data? Læseren, snarere end selve tegnene. En katalogpost er metadata om en bog, men for kataloget er den selve datasættet; et fotos optagelsestidspunkt er en detalje for den, der kigger på billedet, og råmateriale for den, der kortlægger, hvor billeder bliver taget. Metadata kontra strukturerede data er et snævrere spørgsmål, og svaret er, at strukturerede data er metadata skrevet i et fælles ordforråd, så et program kan læse et felt uden at gætte på, hvad det betyder.

Brug begrebet i en beslutning

Brug »Hvad er metadata?« i en konkret beslutning. Vi kan knytte begrebet til det system, I bygger eller køber. Regler og kontroller, der fanger dårlige data på vej ind, og et klart svar på, hvem der ejer hvilket tal, og hvor det kommer fra.

Se vores ydelser Kontakt os →

Undersøg antagelserne bag ordet

Brug forklaringen og det tilknyttede arbejde til at formulere spørgsmål til jeres projekt. Vi kan undersøge antagelserne sammen, før et værktøj vælges. Én infrastrukturplan til 14 afdelinger — 14 forskningsafdelinger på én infrastrukturplan

Én infrastrukturplan til 14 afdelinger Kontakt os →

Hvorfor det betyder noget

Hvert felt er et lille løfte, og den beslutning, der tæller, er, hvor hvert enkelt henter sin værdi: fra en person, der skriver den, fra versionshistorikken, fra en databasekolonne eller fra det, filsystemet tilfældigvis melder. Programmer handler på de løfter uden at efterprøve dem. Det er hele pointen med metadata, og det er også risikoen.

Caches og crawlere fordeler deres arbejde efter dem. En browser med en kopi af en side sender den `Last-Modified`-dato tilbage, den fik, og når siden er uændret, får den et kort `304 Not Modified` i stedet for hele siden igen. RFC 9110 beder en server sende den dato, når den seneste ændring kan fastslås rimeligt og konsekvent, og betingelsen virker begge veje: hvor der ikke findes en dato, der kan fastslås pålideligt, beder standarden heller ikke om en. Ændringsdatoen i et sitemap gør det samme for en crawler, der skal vælge, hvad den henter igen.

Mennesker og hjælpemidler læner sig også op ad dem. Attributten `lang` fortæller en skærmlæser, hvilke udtaleregler den skal bruge, og titlen er det, en fane, et bogmærke og et søgeresultat viser. Rettigheder rejser på samme måde: en pakkehåndtering kender en afhængigheds licens, fordi afhængigheden selv har erklæret den, så en licensgennemgang bygget på den liste er præcis så god som erklæringerne nedenunder.

Hvert felt skal blive ved med at være sandt, så en side med fyrre felter har fyrre chancer for at tage fejl. Nogle felter læser ingen: HTML-standarden bemærker, at mange søgemaskiner ser bort fra navnet `keywords`, fordi det har været brugt til at spamme søgeresultater. Og metadata lækker, for et fotos position eller et dokuments revisionshistorik kan sige mere end selve indholdet. Få felter, der hver især udfyldes fra en kilde, der ikke kommer ud af trit med indholdet, slår som regel mange.

Sådan ser det ud i praksis

Metadata dukker op i tre af projekterne på dette site, hver gang i en anden slags fil: et sitemap, en række PDF'er og en afhængighedsgraf.

Et sitemap påstod engang, at 172 af 176 sider var ændret i samme øjeblik. Datoen kom fra filernes ændringstidspunkt, og indholdseksporten skriver alle filer om ved hver kørsel. Nu går datoen gennem en kæde af kilder: først et eksplicit felt, så commit-historikken og til sidst filen. Kæden har en fælde, som projektet skrev ned: feltets eget navn skal stå på listen, ellers læser generatoren aldrig feltet. Databasen bag siderne holder desuden styr på, hvornår en post blev skrevet og revideret, adskilt fra de år, hvor selve arbejdet fandt sted, og hvor ingen kender datoen, står feltet blankt, fordi et præcist, men forkert tal bliver troet, mens et manglende tal ikke gør. På det dokumenterede implementeringstrin berørte en gensynkronisering efter en måneds rettelser 8 af 186 filer.

På det dokumenterede valideringstrin bestod ni genererede PDF-dokumenter, der blev målt mod tilgængelighedsstandarden alle 106 konfigurerede regler i tilgængelighedsprofilen. Arbejdet omfattede et erklæret dokumentsprog og en rigtig titel i metadataene sammen med tagget struktur, alternativ tekst og en eksplicit læserækkefølge. I den dokumenterede kørsel fejlede de samme dokumenter på præcis én regel af 146 i arkivprofilen. Projektsiden beskriver, hvor tæt resultatet var på at opfylde kravene, og navngiver den regel, der fejlede, fordi 145 bestående regler ikke er ensbetydende med overholdelse af standarden.

Og da licenserne i 644 Rust-crates blev tjekket ved hvert build, var revisionens erklærede grænse en grænse ved metadata. Tjekket læser hver crates erklærede licens, erklæringer kan være forkerte eller ufuldstændige, og tjekket beviser intet om en crate, der erklærer sig forkert. Det blev to gange med vilje provokeret til at stoppe buildet, hvilket viser, at det virker; om erklæringerne, det læser, er ærlige, kan intet tjek vise.

Hvor det går galt

Fejlene herunder har ét træk til fælles: hvert tjek, der ikke leder efter dem, består.

Hver side melder om en ændring efter hver udrulning. En webserver, der sender en statisk fils ændringstidspunkt som `Last-Modified`, eller en generator, der daterer et sitemap på samme måde, melder om det øjeblik, filen blev kopieret, og en kloning af repositoryet, en eksport eller et container-build kopierer dem alle. Caches revaliderer alt, og crawlere lærer intet. Tag datoer fra versionshistorikken eller fra et felt, et menneske har skrevet, og lad en dato, som ingen kender, forblive fraværende. I Hugo er kilderne en liste, der læses i rækkefølge, og feltets eget navn skal stå i den:

```
enableGitInfo = true

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

En titel eller en linkforhåndsvisning viser forvanskede tegn. Tegnkodningen er metadata om hver eneste anden byte, og en parser, der møder tekst før erklæringen, må gætte. HTML-standarden kræver, at det element, der erklærer den, står helt inden for de første 1.024 byte, og står det først i `<head>`, er der intet at gætte på. På dette site læste Bings URL Inspection engang titlen og begge Open Graph-strenge som Latin-1, fordi en lang tankestreg i titlen kom før `charset`-erklæringen. HTTP-headeren nævnte også UTF-8, men et værktøj, der læser filen uden headeren, har kun siden selv at gå efter. Nu er `charset` det første element, og et tjek kontrollerer rækkefølgen på hver bygget side.

Det samme faktum siger forskellige ting forskellige steder. En sides titel bor i `<title>`, i `og:title`, i dens strukturerede data og, for en PDF, i dokumentegenskaberne; dens sprog bor i `lang`, i `Content-Language` og i `og:locale`. Skrives de hver for sig, glider de fra hinanden, og hver modtager tror på sin egen udgave. Beregn hver værdi ét sted, i én skabelon eller funktion, og lad alle steder, der skriver den ud, læse netop den værdi.

En fil fortæller mere, end forfatteren ville offentliggøre. Fotos bærer kameraet, tidspunktet og ofte positionen; kontordokumenter bærer forfatternavne, redigeringshistorik og skabelonstier; et `generator`-tag navngiver

den software, der byggede en side. Fjern ved eksport det, som ingen har valgt at offentliggøre, og behandl hvert felt som tekst, en fremmed vil læse.

Læs mere

- RFC 9110, section 6.3: Header Fields — hvordan HTTP bærer metadata foran indholdet, og hvad et headerfelt må gøre.
- RFC 9110, section 8.8.2: Last-Modified — hvornår en server bør sende en ændringsdato, og hvorfor en dato i fremtiden erstattes.
- HTML Standard: the meta element — standardnavnene for metadata og reglerne for at erklære en tegnkodning.
- DCMI Metadata Terms — Dublin Cores femten elementer og det ordforråd, der er bygget omkring dem.
- The Open Graph protocol — de fire egenskaber, protokollen kræver på hver side, og de valgfrie omkring dem.

Ordliste

Vælg ud fra jeres faktiske vilkår

Medbring et repræsentativt datasæt, de spørgsmål det skal besvare, og den aktuelle flaskehals. Vi kan aftale målinger af kvalitet, gennemløb og gendannelse.

Se, hvad det koster Kontakt os →

Gør begreberne praktiske

Spørg, hvordan »Hvad er metadata?« gælder i jeres organisation. Et konkret eksempel fra jeres nuværende system er nok til at begynde samtalen.

Se vores ydelser

welcome@engineer.company +45 50 10 61 84 København, Danmark

Tag denne side med dig

Visitkort

Lavet med kærlighed af D.

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

Copyright © 2025 – nu · Engineer ApS · Alle rettigheder forbeholdes.

Udskrevet fra https://engineer.company/da/glossary/metadata/