Engineer

Metadata

data som beskriver andra data, så att de kan hittas

Vad det är

Metadata är data som handlar om andra data: titeln, datumet, språket, upphovspersonen, formatet och storleken som följer med ett dokument, så att människor och program kan sortera det, lita på det och agera på det innan de har läst det. Äldre standarder skriver ordet med bindestreck, meta-data, och i vardagen går det under smalare namn: huvudfält i HTTP, egenskaper i en kontorsfil, taggar i ett foto och front matter överst på en sida skriven i Markdown.

En vanlig indelning har tre sorter. Beskrivande metadata säger vad något handlar om: titel, upphovsperson, ämne, sammanfattning. Strukturell metadata säger hur delarna hänger ihop: kapitel, sidor, läsordning. Administrativ metadata säger var något kommer ifrån och vad man får göra med det: när det skapades och ändrades, av vem, under vilken licens och i vilket format. Dublin Core, femton element som blev standarden ISO 15836 år 2003, är den mest kända gemensamma listan över beskrivande fält, och dess titel, upphovsperson, datum, språk och rättigheter återkommer under andra namn i nästan varje format som kommit sedan dess.

En webbsida bär metadata i lager. Först kommer svaret i HTTP. Enligt RFC 9110 kan ett huvudfält ändra eller utvidga ett meddelandes betydelse, beskriva avsändaren, definiera innehållet eller ge ytterligare sammanhang, och `Content-Type`, `Content-Language` och `Last-Modified` beskriver sidan innan en enda byte av den har kommit fram. Inne i HTML-koden finns `<title>`, attributet `lang` och `<meta>`-element vars namn HTML-standarden definierar, t.ex. `description`, `author` och `generator`. Ovanpå dem ligger vokabulär som andra parter läser: egenskaperna i Open Graph för förhandsvisningar av länkar, och strukturerad data för sökmotorer och assistenter.

Vad skiljer då metadata från data? Läsaren, snarare än bytena. En katalogpost är metadata om en bok och hela datamängden för katalogen själv. Tidpunkten då ett foto togs är en detalj för den som tittar på bilden och råmaterial för den som kartlägger var foton tas. Hur metadata förhåller sig till strukturerad data är den smalare frågan, och svaret är att strukturerad data är metadata skriven i en gemensam vokabulär, så att ett program kan läsa ett fält utan att gissa vad det betyder.

Använd begreppet i ett beslut

Använd ”Vad är metadata?” i ett konkret beslut. Vi kan knyta begreppet till systemet ni bygger eller köper. Regler och kontroller som fångar dålig data på vägen in, och ett tydligt svar på vem som äger vilken siffra och var den kommer ifrån.

Se våra tjänster Hör av er →

Granska antagandena bakom ordet

Använd förklaringen och det länkade arbetet för att formulera frågor till projektet. Vi kan undersöka antagandena tillsammans innan ett verktyg väljs. En infrastrukturplan för 14 avdelningar — 14 forskningsavdelningar under en och samma infrastrukturplan

En infrastrukturplan för 14 avdelningar Hör av er →

Varför det spelar roll

Varje fält är ett litet löfte, och beslutet som räknas är varifrån varje fält får sitt värde: från en person som skriver in det, från versionshistoriken, från en kolumn i en databas eller från det filsystemet råkar rapportera. Program agerar på de löftena utan att kontrollera dem. Det är hela poängen med metadata, och också risken.

Cachar och sökrobotar fördelar sin möda efter den. En webbläsare som har en kopia av en sida skickar tillbaka det `Last-Modified` -datum den fick, och om ingenting har ändrats får den ett kort `304 Not Modified` i stället för sidan på nytt. RFC 9110 ber en server skicka det datumet för varje representation vars senaste ändring rimligen och konsekvent går att fastställa, och villkoret skär åt båda hållen: där datumet inte kan fastställas på ett tillförlitligt sätt ber standarden inte om något. Ändringsdatumet i en webbplatskarta gör samma jobb för en sökrobot som väljer vad den ska hämta igen.

Människor och hjälpmedel är också beroende av den. Attributet `lang` talar om för en skärmläsare vilka uttalsregler den ska använda, och titeln är det som en flik, ett bokmärke och en träff i en sökmotor visar. Rättigheter färdas på samma sätt. En pakethanterare känner till licensen för ett beroende bara för att beroendet har deklarerat den, så en licensgranskning som bygger på den listan är exakt så bra som deklarationerna under den.

Varje fält måste förbli sant, så en sida med fyrtio fält har fyrtio chanser att ha fel. Vissa fält läses av ingen alls: HTML-standarden noterar att många sökmotorer bortser från namnet `keywords`, eftersom det användes för att spamma träfflistor. Och metadata läcker, eftersom positionen i ett foto eller revisionshistoriken i ett dokument kan säga mer än innehållet. Färre fält, vart och ett fyllt från en källa som inte kan glida isär, slår oftast många.

Så ser det ut i praktiken

Metadata dyker upp i tre projekt på den här webbplatsen, varje gång i en annan sorts fil: en webbplatskarta, en uppsättning PDF-dokument och en beroendegraf.

Ändringsdatumen i en webbplatskarta är metadata om sidorna, och under en period uppgav kartan samma ändringsögonblick för 172 av 176 sidor. Fältet fylldes från filernas ändringstider, och innehållsexporten skapar alla filer på nytt när den körs, så en enda synkronisering satte ny stämpel på hela innehållet. Åtgärden låter generatorn leta i en fast ordning, först i ett fält som författaren anger, därefter i commit-historiken och sist i filen, och den dokumenterar en fälla i den ordningen: om fältets bokstavliga namn saknas i listan läses fältet aldrig. Posterna bakom sidorna håller nu isär den dag varje post skrevs och reviderades från de år arbetet utfördes, och ett datum som ingen känner till lämnas tomt, enligt principen att ett exakt men felaktigt tal är värre än ett som saknas. Med datumen hämtade ur historiken berörde en synkronisering efter en månads ändringar 8 filer av 186 i det dokumenterade implementeringsskedet.

Vid det dokumenterade valideringsskedet klarade nio genererade PDF-dokument som mättes mot tillgänglighetsstandarden alla 106 konfigurerade regler i tillgänglighetsprofilen. Arbetet omfattade ett deklarerat dokumentspråk och en riktig titel i metadata, tillsammans med taggad struktur, alternativtext och en explicit läsordning. I den dokumenterade körningen underkändes samma dokument på exakt en regel av 146 i arkivprofilen. Sidan redovisar hur nära resultatet kom att uppfylla kraven och namnger den underkända regeln, eftersom 145 godkända regler inte innebär överensstämmelse med standarden.

Och när licenserna för 644 Rust-crates kontrollerades vid varje bygge var granskningens uttalade gräns en gräns hos metadata. Den läser varje licens som en crate har deklarerat, deklarationer kan vara felaktiga eller ofullständiga, och kontrollen bevisar ingenting om en crate som deklarerar sig fel. Den fick fallera med flit två gånger, vilket visar att kontrollen fungerar. Vad en kontroll än gör kan den inte visa att deklarationerna den läser är ärliga.

Var det går fel

Felen nedan har ett drag gemensamt: varje kontroll som inte letar efter dem går igenom.

Varje sida rapporteras som ändrad efter varje driftsättning. En webbserver som skickar en statisk fils ändringstid som `Last-Modified`, eller en generator som daterar en webbplatskarta på samma sätt, rapporterar ögonblicket då filen kopierades, och en utcheckning, en export eller ett containerbygge kopierar dem allihop. Cachar validerar om allt och sökrobotar lär sig ingenting. Ta datumen från versionshistoriken eller från ett fält som en person har skrivit, och låt ett datum som ingen känner till förbli tomt. I Hugo är källorna en lista som läses i ordning, och fältets eget namn måste stå i den:

```
enableGitInfo = true

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

En titel eller en länkförhandsvisning visar trasiga tecken. Teckenkodningen är metadata om varje annan byte, och en parser som möter text före deklarationen måste gissa. HTML-standarden kräver att elementet som deklarerar den ryms helt inom de första 1 024 bytena, och att lägga det först i `<head>` lämnar ingenting att gissa. På den här webbplatsen läste URL Inspection i Bing en gång titeln och båda strängarna för Open Graph som Latin-1, eftersom ett långt tankstreck i titeln kom före teckenkodningen. HTTP-huvudet angav också UTF-8, men ett verktyg som läser bytena utan huvudet har bara sidan att gå på. Teckenkodningen är nu det första elementet, och en kontroll bekräftar den ordningen på varje byggd sida.

Samma uppgift säger olika saker på olika ställen. Titeln på en sida finns i `<title>`, i `og:title`, i dess strukturerade data och, för en PDF, i dokumentegenskaperna. Språket finns i `lang`, i `Content-Language` och i `og:locale`. När de skrivs in var för sig glider de isär, och varje konsument tror på en annan. Räkna fram varje värde en gång, i en mall eller en funktion, och låt varje ställe som skriver ut det läsa just det värdet.

En fil säger mer än författaren tänkte publicera. Foton bär kameran, tiden och ofta positionen. Kontorsdokument bär författarnamn, redigeringshistorik och sökvägar till mallar. En `generator` -tagg namnger mjukvaran som byggde en sida. Rensa vid exporten bort allt som ingen har valt att publicera, och behandla varje fält som text som en främling kommer att läsa.

Läs mer

- RFC 9110, section 6.3: Header Fields — hur HTTP bär metadata före innehållet, och vad ett huvudfält får göra.
- RFC 9110, section 8.8.2: Last-Modified — när en server bör skicka ett ändringsdatum, och varför ett datum i framtiden byts ut.
- HTML Standard: the meta element — standardnamnen för metadata och reglerna för att deklarera en teckenkodning.
- DCMI Metadata Terms — de femton elementen i Dublin Core och vokabulären som byggts runt dem.
- The Open Graph protocol — de fyra egenskaper som krävs på varje sida, och de valfria runt omkring dem.

Ordlista

Välj utifrån era faktiska förutsättningar

Ta med ett representativt dataset, frågorna det ska besvara och dagens flaskhals. Vi kan avtala mått på kvalitet, genomströmning och återställning.

Se vad det kostar Hör av er →

Gör begreppen praktiska

Fråga hur ”Vad är metadata?” gäller i er organisation. Ett konkret exempel från ert nuvarande system räcker för att börja samtalet.

Se våra tjänster

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

Ta med dig den här sidan

Visitkort

Omsorgsfullt skapad med kärlek av D.

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

Upphovsrätt © 2025 – nu · Engineer ApS · Alla rättigheter förbehållna.

Utskrivet från https://engineer.company/sv/glossary/metadata/