SND
Svensk nationell datatjänst

Vad är metadata?

Pass 3: Metadata

BAS Online 2021-01-20

I den här presentationen kommer jag ge en introduktion till metadata och forskningsdata på ett principiellt plan. Vi kommer bland annat titta lite närmare på vad metadata är för något och tar upp några olika definitioner som finns. Det är inte meningen att du ska kunna alla detaljer, utan att du ska kunna visa på poängen med metadata för forskningsdata och kunna prata om det här med forskare. Jag kommer också visa lite på vilka skillnader som finns mellan data, metadata och dokumentation.

Vad tänker du på när du hör ordet metadata? Du kanske har egna erfarenheter av att arbeta med metadata i någon form, och tänker på katalogisering av publikationer, arkivering av handlingar, eller kanske något helt annat.

För att kunna säga något om vad metadata är, kan det vara en god idé att först titta på vad forskningsdata egentligen är. Vi har redan tagit upp det tidigare i kursen, men här kommer ändå en liten uppfräschning av minnet.

Synen på vad som är forskningsdata skiljer sig åt mellan olika institutioner och ämnesområden, men också mellan olika myndigheter och organisationer. Forskningsdata är det material som samlats in under forskningsprocessen för att tjäna som underlag till olika vetenskapliga analyser. Det kan exempelvis vara i form av siffror, databaser och text i olika former, men materialet kan också bestå av biologiska prover, fysiska samlingar, programkod, geodata och mycket, mycket mer.

Som du ser finns det många olika typer av forskningsdata. Ibland förekommer de i fysisk form, till exempel som pappersenkäter, men ofta digitaliseras de för att underlätta hantering och förvaring. En enkel och sammanfattande definition av forskningsdata som vi använder vid SND är att forskningsdata är något som analyseras i ett vetenskapligt syfte – för att besvara en vetenskaplig frågeställning. Vilka typer av data som det handlar om kan därför variera kraftigt.

Så om vi återgår till det här med metadata. Vad är det för något? Det mest banala svaret är att metadata är ett ord med 8 bokstäver. Antalet bokstäver är då metadata om själva ordet ”metadata”.

Metadata kan definieras på olika sätt. I boken *Metadata* av Jeffrey Pomerantz som du hittar i kursens litteraturlista så definieras metadata som ”a statement about a potentially informative object”. I det här sammanhanget motsvarar det potentiellt informativa objektet forskningsdata. Metadata är alltså nånting som säger nånting om forskningsdata. En något mer rättfram definition kommer från National Information Standards Organization i USA. De säger att metadata är ”structured information that describes, explains, locates, or otherwise represents something else”[1]. I det här fallet är ”something else” lika med forskningsdata. Ytterligare en definition som ibland används är att metadata är ”the sum total of what one can say at a given moment about any information object at any level of aggregation”[2]. Det här är en definition som fångar in allt som beskriver objektet i fråga och på alla nivåer. Ett mer klatschigt sätt att sammanfatta de här ganska krångliga definitionerna är att säga att metadata är data om data. Det är också den vanligaste och enklaste definitionen man stöter på.

Vi ska nu titta på begreppen data och metadata i förhållande till vad som är vad. När man pratar med forskare är det viktigt att tänka på att vi, beroende på tidigare erfarenheter, kan ha olika föreställningar om vad som är det ena och vad som är det andra.

Vid en första anblick kan det verka som att det är ganska lätt att hålla isär vad som är data och vad som är metadata. Det ena är ju det material som samlas in och analyseras av en forskare och det andra är information om det insamlade materialet. Fast om man tänker på uppdraget att tillgängliggöra forskningsdata för ny forskning, så kan ju data som samlats in för ett syfte återanvändas av en annan forskare för ett nytt syfte. Det innebär att det som är data för den ene mycket väl kan vara metadata för den andre. Vad som är vad beror på forskningsfrågan och vad man har definierat som analysenhet för sin forskning.

Ett exempel på det här är Språkbanken. Språkbanken är en forskningsinfrastruktur som samlar in enorma mängder text till så kallade korpusar, som de publicerar på sin webbplats. I Språkbankens korpussökverktyg, kan man göra ganska avancerade språkvetenskapliga sökningar. Till exempel går det att söka på alla förekomster av ett visst substantiv i bestämd form. I det här exemplet tittar vi på ordet ”skolan”. I en mening från tidningen Åbo Underrättelser 2012 kan man läsa: ”skolan renoveras, personal anställs men den slutar kort därefter eftersom lönen aldrig betalas ut”. Språkbanken ger oss exempel på var det här förekommer i olika tidningar och andra texter som de har hämtat från internet och diverse andra källor. De här sökningarna bygger på att det finns information om att det är bestämd form på ett visst ställe. Så alla de här texterna har blivit uppmärkta med information om varje ord som talar om att just det här ordet, ”skolan” i det här fallet, är ett substantiv, det är bestämd form, det är singularis och det har en viss funktion i meningen, till exempel är det ett subjekt. Det finns också länkar till ett lexikon som säger att ordformen i fråga hör till ett visst uppslagsord. I lexikonet finns det mycket information om vartenda ord i de här textmängderna. All den här uppmärkningen som berättar allt runt omkring ordet kan tyckas vara typisk metadata eftersom det säger något om det här ordet, som i sig utgör data.

Så är fallet för någon som forskar om byggnadsvård och från den här meningen kan få reda på att en viss skola har renoverats. Så är också fallet för någon som forskar om personalen eller personalpolitiken på skolan i fråga. MEN för en forskare som är intresserad av grammatik är kanske själva ordet tämligen ointressant och det är informationen om att det är ett substantiv i singularis som är subjekt till ett passivt verb som är det intressanta och som är data.

Så vad som är data och metadata beror liksom på. Är man intresserad av innehållet i texterna är uppmärkningen metadata. Men om man är intresserad av formen är det snarare uppmärkningen som faktiskt är data. Av det här exemplet förstår vi att vad begreppen data och metadata betyder beror på vilken forskning som bedrivs och att det är viktigt att vara medveten om vad som är vad för att undvika missförstånd och för att ha en gemensam terminologi när man pratar metadata med forskare.

Ett annat begrepp som kan vara bra att reflektera över är dokumentation. I många fall används metadata och dokumentation som synonyma begrepp. Fast det finns faktiskt vissa skillnader. För en forskare kan dokumentation ofta vara ett mer lättförståeligt begrepp än metadata, för man har kanske hört betydligt mindre om metadata. I den mån det ändå finns en skillnad mellan dokumentation och metadata kan man säga att metadata är strukturerade för att kunna läsas av både människor och datorer. Ett sätt att göra det på är att använda XML-scheman som består av standardiserade termer.

Dokumentation behöver däremot inte vara maskinläsbar och har inte alls samma krav på struktur. Exempel på dokumentation kan vara exempelvis forskningsrapporter, artiklar och metodbeskrivningar.

Sammanfattning

I den här presentationen har jag börjat med att introducera begreppet metadata. Vi har gått igenom olika exempel på definitioner, där det enklaste sättet är att säga att metadata är data om data. Vi har också problematiserat gränsdragningen mellan vad som är data, vad som är metadata och vad som är dokumentation och att man behöver känna till att det som är metadata för någon kanske är data för en annan, och kanske dokumentation för en tredje.

Referenser

[1] https://www.niso.org/publications/understanding-metadata-2017

[2] http://www.getty.edu/publications/intrometadata/setting-the-stage/