CIDOCs arbeidsgruppe for koreferanse

av Øyvind Eide (leder for arbeidsgruppa), Enhet for Digital Dokumentasjon, Universitetet i Oslo

Innledning

Tenk deg at du er opptatt av grensearbeidet som ledet fram mot grenseavtalen mellom Danmark-Norge og Sverige-Finland i 1751. Det finnes en rekke kilder til dette arbeidet, og i disse kildene omtales mange historiske personer. Et av de mange spørsmålene man kan ønske seg svar på er om det finnes vitner som både ble forhørt av Major Schnitler på norsk side og av de svenskene grenseingeniørene. Dersom man finner slike personer, er det spennende å se hva de sa til representantene for de to ulike autoritetene.

For å finne ut av dette, må man identifisere personene som er omtalt ved navn i de ulike kildene, og finne ut om noen av dem går igjen i flere kilder. Det er ikke alltid mulig å identifisere alle personer som er omtalt i en tekst, men i dette tilfellet kan vi identifisere mange av dem. Litt mer abstrakt innebærer dette at man må finne ut hvilke historiske personer de ulike navnene refererer til, og om det finnes referanser til samme historiske person i flere kilder. Dette er referanser fra tekst til historisk virkelighet, og koreferanser fra flere ulike tekster til samme historiske person.

Koreferanse er således basert på begrepet referanse. I eksemplet over snakket vi om tekstfragmenter som referer til historiske personer, men det kunne vært alle slags fysiske objekter så som personer, steder eller objekter, og også abstrakte objekter. Koreferanse oppstår når to eller flere tekstfragmenter peker på det samme objektet i verden. Et annet eksempel kan være et avsnitt i en arkeologiske rapport og en innførsel i en museumskatalog som omtaler samme gjenstand.

Koreferanser innad i en organisasjon og innad i ett enkelt informasjonssystem er viktig, men ikke det sentrale for vårt arbeid. I arbeidsgruppa konsentrerer vi oss om koreferanser som finnes mellom samlinger i ulike institusjoner.

Tradisjonelle metoder

Forskere, konservatorer, bibliotekarer og studenter bruker mye tid på å finne koreferanser. Omhandler denne teksten samme person som vi ser på dette bildet? I slikt arbeid er autoritetsregistre et viktig verktøy, som hjelper oss å finne koreferanser.

Å finne koreferanser er svært tidkrevende arbeid. Dessverre tas ikke resultatene av arbeidet godt nok vare på. Opplysninger om koreferanser finnes i mange tilfelle kun i hodet på den som oppdager dem. Ofte har det blitt skrevet ned som et notat, kanskje på et arkivkort, og i noen tilfelle har slike opplysninger kommet med i publikasjoner, f.eks. i form av fotnoter.

Felles for alle disse resultatene er at de er vanskelig tilgjengelige for andre, og at de ikke egner seg for bruk i datasystemer. Det betyr at vi trenger bedre metoder for å uttrykke denne informasjonen i en digital tid. Den bør uttrykkes maskinleselig på en slik måte at den kan bruke til andre ting enn å leses av mennesker. I tillegg bør man ta vare på viktig informasjon om hvor opplysningen om koreferansen kommer fra: Hvem er ansvarlig, når skjedde det, hva er kildene, osv. Dette er viktig for å gjøre opplysninger i museer og andre kulturarvinstitusjoner mer etterprøvbare.

Koreferanser i en digital tid

Opplysninger om koreferanser bør derfor lagres i en eller annen form for formalisert oppsett. Det kan for eksempel være en databasetabell eller et XML-basert system. Det viktige er at man lagrer en eksakt referanse til de to stedene koreferansene finnes, og grunnlaget for påstanden om koreferanse.

Dette kan gjøres på flere måter i praksis. Dels kan man tenke seg at opplysninger om koreferanser lagres i datasystemene der referansene er lagret. Ved Enhet for Digital Dokumentasjon, Universitetet i Oslo er vi i ferd med å utvikle et slikt system, der alle opplysninger i databasene våre kan korefereres mot opplysninger alle andre steder, være seg i databaser, i web-systemer eller i trykte bøker eller andre papirbaserte systemer.

En annen måte å gjøre dette på er å lage et system for å lagre koreferanser mellom opplysninger i to ulike eksterne kilder. Information Systems Lab ved FORTH-ICS i Hellas er i ferd med å utvikle et system der man kan lagre koreferanser mellom to ulike web-adresser. Arbeidsgruppa for koreferanse i CIDOC vil vi i løpet av neste år utvikle felles dataformater for å kommunisere slike opplysninger mellom ulike institusjoner og mellom ulike datasystemer.

Identitetsnettverket

Slike systemer er nyttige for de enkelte institusjoner, og det er god nok grunn i seg selv til å utvikle dem. Men i tillegg finnes det spennende muligheter til å forske på selve koreferansene som informasjonssystem. Når man har opplysninger om koreferanser fra en rekke ulike institusjoner, og disse opplysningene er uttrykt i et felles formelt språk, kan man bygge opp et identitetsnettverk. Ved hjelp av dette kan man undersøke relasjoner mellom personer.

Sammen med flere kollegaer arbeider Carlo Meghinis ved CNR–ISTI i Pisa, Italia med en formell beskrivelse av hvordan slike systemer kan bygges opp på en konsistent og fungerende måte. Dersom man utvikler slike systemer og de fylles opp av opplysninger, vil man kunne stille spørsmål av typen som ble skissert i eksempelet fra 1700-tallets grensearbeid ovenfor og få meningsfulle svar. Dette gjør at man kan gjennomføre innsamling av informasjon raskere, og komme fram til det mest spennende: Å tolke opplysningene og derigjennom finne ut mer om fortiden.

Referanser

- CIDOCs arbeidsgruppe for koreferanse: http://cidoc.mediahost.org/co_reference_wg(en)(E1).xml
- Beskrivelse av Enhet for Digital Dokumentasjons system for koreferanser knyttet til databasene (under utvikling): http://cidoc.mediahost.org/eddSystemCoref.pdf
- Beskrivelse av Information Systems Lab ved FORTH-ICS i Hellas' prototype for et system for lagring av koreferanser på web: http://cidoc.mediahost.org/Tagging-Tool.pdf
- Carlo Meghini, Martin Doerr, Nicolas Spyratos: "Managing co-reference knowledge for data integration." S. 229-248 i: Proceedings of EJC2008, the 18th European-Japanese Conference on Information Modelling and Knowledge Bases. Tsukuba, Japan, 2008.