Digitalt skapt kildemateriale og arkivarenes påvirkning - fast element i kildekritikken?

Av

Børge Strand

Forskerforum

- Fra 2014 har det vært en rekke artikler og debattinnlegg om
  - Registerbasert forskning
    - 1/2014,s 35, 9/2014,s 42
    - verdens beste dataregistre
  - Bekymring for tilgangen til mikrodata for forskning
- SSB har et tilnærmet monopol på 'utlån av mikrodata for forskning'
- SSBs egne registre og eksterne, administrative registre
  - Tilrettelegger -'skreddersyr' - datagrunnlag for forskere
  - Prisøkning på tilrettelegging, SSBs tiltak(2/2015, s 37)
- Dette er i stor grad de samme registrene som arkivinstitusjonene bevarer

Forskningens etterspørsel og bruk

- Stor og økende etterspørsel etter mikrodata fra registre – alltid informasjonsverdien som etterspørres
- SSB tilrettelegger og distribuerer mikrodata til forskere – avidentifisert eller anonymisert
- Ingenting av dette kommer fra registre/fagsystemer som er 'godkjent' av Riksarkivaren
- Ingenting har vært gjennom arkivdanning detaljregulert av Arkivverket
- Ingenting av dette datagrunnlaget har vært innom arkivinstitusjoner og fått et stempel som 'autentisk'
- Ingenting av dette datagrunnlaget er forseglet med sjekksummer
- Likevel brukes det – og tilliten til registermaterialet er stor
- SSB og forskerne selv kvalitetssikrer datagrunnlaget og vurderer beviskraften, feilmarginer etc. - øver kildekritikk

Registre/fagsystemer

- En gullgruve for forskning – historieforskning og annen forskning
- Informasjon i strukturert form: I tabeller, rader, kolonner og celler
- Tilrettelagt for datautveksling og gjenbruk ved kobling – maskinell bruk
- Informasjon i kodet form – i stor grad tallkoder (yrke, utdanning, bosted, sivilstand …)
- Dette er egenskaper som gir store gevinster i forskningssammenheng
- Alt er klausulert og må avidentifiseres eller anonymiseres før utlevering til forskeren
- Lagres som 'rådata' i arkivdepot (og hos SSB) – dvs. vi mister opprinnelig funksjonalitet for spørring, gjenfinning, prosessering
- Men som forskere vil vi uansett stille nye og andre spørsmål til materialet - ikke gjenta arkivskapers spørringer og bruk

Informasjonsverdien i registre

- Egnet for å avdekke strukturer, mønstre, samvariasjon, endringer over tid, livsløp, geografiske variasjoner m.m.m.
- Også brukes til enkeltoppslag etter mønster fra digitalisert materiale – søk etter en og en person (slektsforskning)
- Hvis enkeltoppslag blir den eneste form for bruk, er det en skrøpelig utnyttelse av registermaterialet
- Forskere etterspør registermateriale ‘skreddersydd’ for et gitt forskningsprosjekt – (DIP) - data om grupper/populasjoner, i tabellform for maskinelle analyser
- Forskere vil ikke søke opp en og en opplysning

Arkivskaperes og publikums etterspørsel

- Skiller seg fra forskningens
- Enkeltoppslag – rettighetsdokumentasjon
- Strengere krav til pålitelighet?
- Dokumentasjon av historisk, pensjonsgivende inntekt – et typisk eksempel på publikumsforespørsel
- Dette er dokumentert i Ligningsregisteret (SKD) – årlige datasett fra 1967
- Katalogen – opprinnelig og ADDML-fil
- Script for søking - eks
- Dataflyt fra Ligningsregisteret til Det Sentrale Folketrygdsystem (NAV)
- Skattelistene er utskrifter fra Ligningsregisteret - Ligningsregisteret er primærkilden både for listene og for DSF
- Et stort antall slike forespørsler – aldri fått tilbakemelding med spørsmål om nærmere dokumentasjon

Registermateriale og kildekritikk

- 'Maskinell kildekritikk'
- Logiske kontroller - sammenhenger og gjensidig avhengighet, systemgenerert informasjon, verdiområder for variable
- Konsistens – samsvar – avvik - internt i registeret
- Validering av fødselsnummer, organisasjonsnummer, kommunenummer
- Referanseintegritet
- Sammenligning på makronivå – datautveksling og gjenbruk
- Sammenligning på mikronivå - 'registervasking' – mellom registre
- Retning av dataflyt
  - hvor kommer data fra?
  - avhengighet mellom kilder?
- Entitet:
  - Person, familie, husholdning, skattyter, organisasjon
  - Adresse – ulike nivåer av numerisk adresse
- Mye avvik kan forklares med ulike entiteter og ulike perioder

‘Leverandører’ av registerdata

- Statistisk sentralbyrå (SSB)
- Norsk Samfunnsvitenskapelig Datatjeneste (NSD)
- Forskningsinstitusjoner
- Arkivverket
- Arkivinstitusjoner (interkommunale/kommunale/andre
- Museer?
- Arkivskaperne selv?

Hver sine 'nisjer'

- SSB er en stor aktør som i mange år har levert mikrodata til forskere – anonymisert eller avidentifisert, men stort sett bare 'ferske' data, dvs. nyeste tilgjengelige
  - samfunnsvitenskapelig forskning
  - prognoser, beslutningsgrunnlag
- Arkivinstitusjonene tilbyr data langs den historiske aksen
- Arkivinstitusjonenes materiale er velegnet for å sette sammen paneldata
- Samme type datagrunnlag, men hver sin 'nisje'

Arkivinstitusjonenes påvirkning

- Fast spørsmål for kildekritkken - hvordan har arkivinstitusjonene påvirket kildematerialet?
- Regelverk og inngripen gjennom livssyklusen
- Hva som kommer inn i arkivet og på hvilken form
  - hvilke tabeller, hvilke felt, hva slags format? Lite rom for arkivskaperne
- BK – hva som kommer inn i depotet – og ikke minst på hva slags form
  - mikrodata som pdf?
- Dette er valg som legger rammer for etterspørselen
- Prosessen med arkivuttrekk - er det uproblematisk at arkivdepotene lager SIP?
- eArkiv og synkron overføring - 'kontinuerlig feilretting'? Arkivinstitusjonen som medaktør i arkivdanningen - en aktiv part og en 'med-arkivskaper'
- Heldigvis har Riksarkivaren holdt fingrene fra fatet i registerverden – så langt
- (Folketellinger, system for ligning, matrikkelen ….)

NOARK-3-uttrekk – eks. på feil og mangler

- Dubletter på primærnøkkel – manglende referanseintegritet
- Manglende løpenummer i saksnr.
- Datofelt som ikke validerer
- Mengder av saker som ikke er avsluttet – skal de avsluttes maskinelt? Hvem skal 'avslutte' dem?
- Mengder av dokumenter som ikke er knyttet til sak – kan disse gjenfinnes?
- Systemene har tillatt dette – til tross for Riksarkivarens godkjenning
- Dette er autentiske feil – 'arkivdanningen er ikke perfekt' - skal arkivdepotene kreve at feilene rettes opp?

Eksempel på forespørsel fra arkivskaper

- Sosialtjenesten lokalt og fagsystemet Humanus (Telenor Allianse)
- Fagsystem som var i bruk 1998 – 2002 – alt innhold er mikrodata i strukturert form – selvsagt KLAUSULERTE data
- Uttrekk til IKA Øst 2014
- Uttrekket produsert av HIKT – tekniske metadata produsert av IKA Øst
- Forespørsel fra NAV i kommunen om utbetalingshistorikk for klient
- Svaret – en tabell – strukturerte data
- Grunnlaget for svaret – sammenstilling av tre tabeller og et antall felt fra hver tabell – fødselsnummer som søkebegrep
- Klienten og utbetalingshistorikken er det viktige her – kommunen kan få tilbake drøyt 200 000 kroner

Humanus – forespørsel forts.

- Kan vi ha tillit til dette? Hvilken beviskraft har det?
- 'Uberørt' av arkivarer – bare arkivskaper og systemleverandør
  - Et system for å håndtere sosialtjenestens behov – flere moduler
    - klientmodul
    - økonomimodul
    - statistikkmodul
    - systemmodul osv.
    - til sammen 75 tabeller – ER-diagram modul 'Klient-økonomi'
  - Systemleverandøren har utviklet systemet i forhold til funksjonen (sosialtjeneste), lovverk og kundenes krav
  - Ikke 'godkjent' av Riksarkivaren – har ingenting med Noark å gjøre
- Kildekritiske vurderinger
  - konsistens (klientdata henger sammen, beløpene samsvarer med satsene, tidspunkter er logiske)
  - utveksler data med andre systemer - lokalt: økonomisystem, personalsystem, med andre kommuners systemer, sentralt: noen ytelser er skattepliktige (FLT)

Bestilling av datagrunnlag for forskning

- Et forskningsprosjekt har utgangspunkt i noen ideer om datagrunnlag – kildemateriale. Hva finnes?
- For registermateriale er det addml-filene som er katalogen – hvor finnes data for mitt konkrete forskningsprosjekt?
- Koblinger – deretter anonymisering!
- Arkivarene må være i stand til å veilede publikum i dette landskapet
- Hvilke arkivskapere, hvilke registre, hvilke tabeller, hvilke felt? Hva slags populasjon?
- Datagrunnlag skreddersydd for et gitt forskningsprosjekt (DIP)
  - Nytt originalmateriale, men kan ikke gjenbrukes i andre sammenhenger
  - For mitt forskningsprosjekt ville bestillingen sett slik ut
- Hvordan ville arkivarene håndtert en slik bestilling?