Om bruk av digitalt skapt arkivmateriale i historieforskning

Av

Børge Strand

Daglig leder IKA Øst

Litt om meg selv og bakgrunn for boken

- Utdanning:
  - Cand. philol. fra Norges teknisk-naturvitenskapelige universitet (NTNU) med historie hovedfag
  - Informatikk fra Norges Informasjonsteknologiske Høgskole (NITH)
- Praksis:
  - Statistisk sentralbyrå (SSB) ca. 12 år
  - Arkivverket (Riksarkivet og Statsarkivet i Hamar) ca. 16 år
  - Fra 2012 daglig leder i Interkommunal arkivordning Øst IKS
- Fellesnevner: 'Mikrodata' - elektroniske, administrative registre som grunnlag for statistikkproduksjon, for forskning og langtidsbevaring av tilsvarende datagrunnlag
- Den aktuelle boken er resultat av et forskningsprosjekt i Arkivverket

Noen stikkord for forskningsprosjektet

- Historieforskning
- Inneholde et case
- Bygge på digitalt skapt arkivmateriale
- Personvern - adgangsbegrensning
- Metodikk som har overføringsverdi
- Målgruppen er historikere, arkivarer, forskere generelt
- Tittelen ga seg selv: 'Digital archives in historical research'
  - Noe overraskende var denne tittelen 'ledig'
  - Mange beslektede titler, men ved nærmere ettersyn: enten digitalisert arkivmateriale eller digitalisert bibliotekmateriale

Hovedgrupper av digitalt arkivmateriale

Systemer for intern administrasjon		Fagsystemer	
Journalføring, sak/arkivsystemer	Andre internadm. systemer (lønn, regnskap, personalmed mer)	Administrative systemer /registre	Forsknings-registre/statistikk-registre

- Registrene er det eldste digitale arkivmaterialet – fra 1960-tallet og fremover
  - men informasjon om enhetene i registrene har lengre historikk
- Informasjonsverdi - dokumentasjonsverdi
- Maskinell – visuell bruk

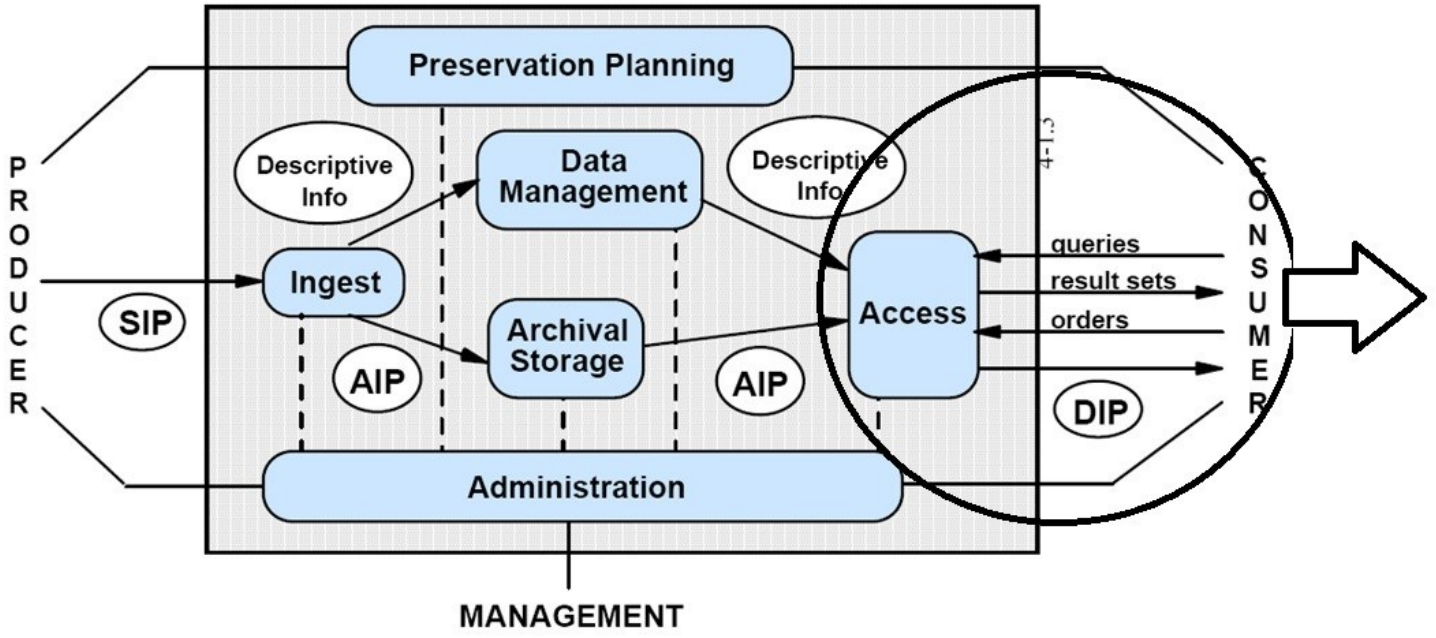
Registre/fagsystemer

- En gullgruve for forskning – historieforskning og annen forskning
- Informasjon i strukturert form: I tabeller, rader , kolonner og celler
- Informasjon i kodet form – først og fremst tallkoder (yrke, utdanning, bosted, sivilstand …)
- Tilrettelagt for datautveksling og gjenbruk ved kobling – maskinell bruk
- Dette er egenskaper som gir store gevinster i forskningssammenheng
- Lagres som 'rådata' i arkivdepot – dvs. vi mister opprinnelig funksjonalitet for spørring, gjenfinning, prosessering
- Men som forskere vil vi uansett stille nye og andre spørsmål til materialet – vi vil bruke dette på helt andre måter

Kobling med individ som kombinasjonsenhet

- Vi vil svært gjerne kombinere informasjon fra ulike systemer og fra ulike arkivskapere – 'kobling'
- Noen nasjonale koblingsnøkler – dvs. entydig identifikasjon av entitetene:
  - Fødselsnummer (tilsv. CPR-nr) og familienummer
    - D-nummer
  - Numerisk adresse
  - Organisasjonsnummer
  - Mange andre ...
- Koblingsnøklene er en egenskap som skiller digitalt skapte registre fundamentalt fra digitaliserte registre
- Kobling er kontroversielt – Datatilsynet - personvern

Plassering i forhold til OAIS



Gangen i forskningsprosjektet

- Problemstilling – teorier – hypoteser - design
- Identifisere, lokalisere, tilrettelegge (evt. bestille) datagrunnlag (DIP)
- Operasjonalisere variable
- Kildekritikk
- Analyse
- Evaluering
  - Av elektronisk arkivmateriale - anvendt i dette konkete tilfellet og mer generelt
  - Hvilke problemer dukket opp underveis?
  - Valg og kompromisser underveis?

Tema og analysemetode

- 'Det postindustrielle arbeidsmarkedet' - kjønn – generasjon - nasjonalt - regionalt - lokalt perspektiv
- Tidsrom ca.1967 - 2007
- Populasjon – alle personer født 1937 - 1958
- Datagrunnlagets art er slik at kvantitativ analyse er naturlig:
  - Korrelasjonsanalyse - avdekker samvariasjon mellom variable
  - Regresjonsanalyse - skiller mellom variabler som forklarer og variabler som blir forklart og mål for styrken på årsakseffekten
  - Kvantitative variable
- Dette er metoder som er styrende og retningsgivende for å identifisere aktuelle datakilder

Fra hypoteser og modellformulering til ferdig datagrunnlag

- Nullhypotese og alternative hypoteser - figur
- Regresjonsanalyse forutsetter en avhengig variabel – det som skal forklares, resultatet (Y)
- En eller flere uavhengige variable (X)– som forklarer resultatet
- Min modell inkluderer i utgangspunktet 5 forklaringsvariable - årsaksdiagram
- En datamatrise som skal inneholde en nærmere definert populasjon…
- Og nærmere definerte variable
  - Analysevariable
  - Teknisk variable, for eksempel for grupperingsformål
  - Variablene må være kvantifiserbare størrelser

Design av analysegrunnlaget

- Valget står mellom:
  - Tverrsnittsdata
  - Tidsseriedata
  - Paneldata
- Mitt valg er paneldata – de samme enhetene (observasjonene) skal følges over en gitt periode, i mitt tilfelle gjennom 20 år
- Enheten er det enkelte individ

Kriterier for populasjonen

- Observasjoner som oppfyller flg. krav:
  - Gyldig fødselsnummer
  - Ingen dubletter
  - Ingen D-nummer
  - Aldersspenn som gjør det mulig å følge hvert individ fra 30 til 50 år innenfor perioden 1967 – 2007
    - Dvs. fødselsår f.o.m. 1937 t.o.m. 1957
  - I live ved 50 års alder
  - Gyldig geografisk tilknytning (kommunenummer) i løpet av perioden

Datagrunnlag

- Fra Arbeids- og velferdsetaten (NAV):
  - Det Sentrale Folketrygdsystem (DSF), 1967 – 2007
- Fra Skattedirektoratet:
  - Ligningsregisteret 1975, 1980, 1985, 1990
- Fra Statistisk sentralbyrå:
  - ”Standard for kommuneklassifisering”, 1974, 1985, 1994 og 2003
- Aktuelt, men ikke anvendt:
  - NAV: AA-registeret (yrke)
  - SSB: Folketellingene 1970 – (yrke og utdanning)
  - SSB/SKD: Folkeregisterdata – (barn og mobilitet)
- Skisse av datagrunnlaget

Et blikk på tekniske metadata – 'katalogen'

"Kataloginformasjon" – tekniske metadata

- Tradisjonell filbeskrivelse
  - DSF – persontabell og inntektstabeller
- Arkivverkets standardiserte metadata: XML-filer
  - Ligningsregisteret
- Hvert felt – i hver tabell - må vurderes i den aktuelle sammenhengen

Koblingsnøkler

- Fødselsnummer
  - 11 siffer, 9. siffer viser personens kjønn
  - de to siste siffer er kontrollsiffer
  - fødselsnummeret kan valideres maskinelt
- Numerisk adresse
  - Tallkode for adresse – hierarkisk struktur - totalt 24 siffer
  - Kun kommunenummer og fylkesnummer er aktuelle for dette tilfellet. Verdiområdet er 01- 20 for fylke og 0101 – 2030 for kommune

Verktøy

- For tilrettelegging av data:
  - SAS – opprinnelig ”Statistical Analyses System”, et akronym forlatt for lenge siden fordi SAS nå har mye mer omfattende funksjonalitet
  - I SAS skiller vi mellom datasteg og prosedyresteg (SAS-nøkkelord er DATA og PROC)
  - Operasjonalisering av variablene må gjøres i DATA steg
- For analysedelen:
  - SPSS – analyseverktøy, bl.a. med prosedyrer for regresjonsanalyse

Kildekritikk

- Ikke tilstrekkelig med overordnet vurdering av kildene
- I stor grad gjennomføres etter at analysegrunnlaget er ferdig etablert
- I stor grad maskinell – mikronivå – eks. validering av fødselsnummer
- Konsistens – samsvar – avvik
- Retning av dataflyt
  - hvor kommer data fra?
  - avhengighet mellom kilder?
- Dekningsgrad
- Validitet
- Representativitet
- Koblingsproblematikk (matching)
    - Hva er en god match?
    - Hva er en dårlig match?

Kildekritikk forts.

- Periode/tidspunkt vs. longitudinelle data
  - sivilstand – antall barn
  - 'Øyeblikksbilde'
  - Definert periode (kalenderår, kvartal, måned …)
- Entitet
  - Person
  - Familie
  - Husholdning
  - Skattyter
- Mye avvik kan forklares med ulike entiteter og ulike perioder

Etter kildekritikken

- Ferdig analysefil (DIP):
  - Individdata
  - Anonymisert - ikke mulig å identifisere enkeltpersoner
  - Inneholder bare tall
  - Grunnlag for maskinell bearbeiding og maskinell analyse (SPSS)
- Nødvendig å justere årsaksdiagrammet
  - Gjenstående X-variabler: kjønn, generasjon, geografisk mobilitet

Analyse og resultater

- Har modellen forklaringskraft? JA!
  - Kjønn er den viktigste forklaringsvariabelen
    - effekten er negativ – som antatt
    - styrken på effekten varierer sterkt mellom kommuner
  - Generasjon har også forklaringskraft
    - effekten er positiv – som antatt
  - Geografisk mobilitet har ikke forklaringskraft
    - Variabelen ikke godt nok operasjonalisert
- Anvendt på ulike geografiske nivåer?
  - Grafisk illustrasjon – temakart fylke og kommune
    - lys blå = kjønn forklarer mye av variasjonen – stor forskjell mellom kvinner og menn mht. grad av deltakelse i arbeidsmarkedet
    - mørk blå = kjønn forklarer lite av variasjonen, liten forskjell mellom menn og kvinner mht. grad av deltakelse i arbeidsmarkedet
  - Kjønn er den viktigste forklaringsvariabelen uansett sted

Evaluering

- En rekke formelle begrensninger – ikke alle ønskede kilder var tilgjengelige
- Problem med brudd i kontinuitet (Ligningsregister)
- Mange valg og kompromisser - generelt mellom variable som har diakrone egenskaper og variable som ikke er diakrone
- Til dels omfattende programmering for å avlede variable
  - særlig Y-variabelen
- Enkelte variable lot seg ikke operasjonalisere som gyldige målebegreper
- Ulike entiteter – person vs. skattyter
- Men tross alt: et datagrunnlag med representative og valide variable, fullt mulig å analysere på lavt geografisk nivå

Overføringsverdi?

- Metoden kan anvendes på registerdata generelt
  - Egnet for å avdekke strukturer, mønstre, endringer over tid og geografiske variasjoner
  - Etterspørsel og bruk? Erfaringen viser at det er stor etterspørsel etter mikrodata
  - Kan selvsagt også brukes til enkeltoppslag etter mønster fra digitalisert materiale – søk etter en og en person
  - Hvis enkeltoppslag blir den eneste form for bruk, er det en skrøpelig utnyttelse av registermaterialet
  - Forskere (historikere og andre) vil – forhåpentligvis – etterspørre registermateriale – 'skreddersydd' som DIP

Arkivinstitusjonene

- Må dokumentere innlevert materiale – 'katalogisere' – tekniske metadata
  - Katalogen må være mye mer detaljert enn for papirbasert materiale
- Tekniske metadata
  - Kan distribueres fritt
  - Tilstrekkelig for å bygge opp et definert analysegrunnlag (kravspesifikasjon for DIP)
  - Tilstrekkelig for å skrive programkode (må ikke være SAS …)
- Må være i stand til å veilede publikum i dette materialet
- Forklare hva som kan finnes hvor, inkl. kunnskap om datautveksling
- Må kunne behandle forespørsler og ordre
- Produsere DIP i hht. konkrete kravspesifikasjoner

Arkivinstitusjonene og forskere

- Ethvert forskningsprosjekt vil kreve sitt særegne, individuelle datagrunnlag – dermed uforutsigbart
- Typisk kombinere data fra mange ulike AIP – fra mange ulike arkivskapere og mange ulike datakilder
- Man kan ikke på forhånd lage en 'standard DIP' for hver AIP
- Forskere vil neppe søke opp en og en opplysning fra en og en database, men i stedet bestille et datagrunnlag 'skreddersydd' for det enkelte forskningsprosjekt
  - En teknisk løsning for enkeltoppslag vil først og fremst vareta tilgang til rettighetsinformasjon for den enkelte borger
- Hvordan ville en arkivinstitusjon behandle en slik ordre – fra en ekstern aktør?
  - Forutsatt at ordren kommer fra en forsker som har krav på tilgang
  - Gjøre tilretteleggingen på vegne av forskeren? I så fall - kreve betaling?
  - Slippe til forskeren på innsiden av sikkerhetssonene? (Jeg satt 'på innsiden' av sikkerhetssonene da jeg jobbet med dette)

Flere aktører – i samme 'marked'?

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

SSB

- 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, beslutninsgrunnlag
- Arkivinstitusjonene tilbyr data langs den historiske aksen
- Arkivinstitusjonenes materiale er velegnet for å sette sammen paneldata
- Samme type datagrunnlag – samme lovverk, men hver vår 'nisje'