Forskningsbiblioteket

Mer om litteratur, publisering og forskningsinfrastruktur

4. juni 2019 AV SONDRE S. ARNESEN

Kunsten å reprodusere

Foto av Chris Barbalis

Hvert år publiseres det en hel haug med forskningsresultater i artikler, monografier og antologier. Som regel presenteres

hovedfunnene i flotte tabeller og grafer, eller som sitatvennlige spissformuleringer – naturligvis avhengig av fagområde. Men hva med rådataene som ligger bak resultatene, er de tilgjengelige for alle som ønsker å gå resultatene nærmere etter i sømmene?

De siste årene har vi sett et økende fokus, i form av policyer hos institusjoner og finansiører, på at rådata også skal gjøres tilgjengelige. Bak denne utviklingen ligger et ideal om «Open Science»; forskningsresultater skal være gjennomsiktige og gjøres åpent tilgjengelige for samfunnet, prosessen skal brettes ut, og konklusjonene kunne diskuteres. Statlig finansiert forskning skal komme samfunnet til gode, da skal også rådataene blottlegges, «så åpent som mulig, så lukket som nødvendig» som det heter i de fleste sagaer på området.

Det finnes naturligvis mange gode argumenter for hvorfor også rådataene skal publiseres; juks og fanteri blir litt vanskeligere, skjønt jeg tviler på om de som ønsker å bedrive svindel har problemer med å manipulere rådata også om det skulle stå på det, det har historien lært oss allerede. I tillegg kan andre forskere (ideelt sett) bygge videre på datamaterialet og sette det inn i nye sammenhenger. Det kan åpne for samarbeid innen spesifikke felt og for brobygging på tvers av forskningsmiljøer.

Så hvordan står det til med kravene til publisering av forskningsdata? Har de bidratt til åpnere vitenskap? Kan man lett gjenskape resultatene man leser i åpne tidsskrift?

De utvalgte

For å svare på noe av dette laget jeg en liste over artikler publisert av HVL-ansatte i åpne tidsskrifter i 2018. Totalen landet på 169 artikler hvor lesetilgangen ikke er strupet av abonnementsavgifter. Artiklene er

såkalt «Gull-Open Access», så har du tilgang til internett kan du finne og lese dem. Jeg sorterte deretter listen min på tidsskrift og tok for meg den første og fineste bolken med titler som våre forskere på en eller annen måte har bidratt til å publisere i. Øynene mine landet på BMC ettersom det også er navnet på et anerkjent sykkelmerke (jeg får ikke betalt for å forske, så sånn blir det).

BMC står for BioMed Central og er en av tungvekterne innen Open Access-publisering. Skal du publisere hos dem så vil artiklene ligge helt åpent og tilgjengelig for alle på nett. Forlaget er eid av moderskipet Springer Nature som har rundt 3000 tidsskrift innen alle mulige fagfelt i porteføljen.

Så hvilke krav til publisering av forskningsdata stiller Springer Nature? Av 3000 tidsskrift er det 1600 som har valgt å benytte seg av en eller annen «type» av forlagets standardiserte datapolicier:

Data policy	Beskrivelse av datapolicy	Anta tidssk
Type 1	Data sharing and data citation is encouraged	62
Type 2	Data sharing and evidence of data sharing encouraged	50
Type 3	**Data sharing encouraged and statements of data availability required**	46
Type 4	Data sharing, evidence of data sharing and peer review of data required	6

Kilder: https://www.springernature.com/gp/authors/research-data-policy/research-data-policy-types og https://www.springernature.com/gp/authors/research-data-policy/journal-policies-and-services

Som vi ser av tabellen er de to første variantene av typen «hyggelig om du publiserer forskningsdataene», og det er også disse majoriteten av tidsskriftene tilbyr sine forfattere. Den fjerde og strengeste varianten er det verdt å merke seg at kun 6 av 3000 tidsskrift har gått for – at data skal publiseres. BMC har for sine tidsskrift valgt å gå for variant 3 (uthevet). Det oppfordres til å dele data og man skal uansett gi en uttalelse om status for tilgang til rådata.

Status for tilgang til forskningsdata

Forskere fra HVL var totalt involvert i 20 artikler i ulike tidsskrift under BMC-paraplyen i 2018. Av disse oppgis det i 7 av tilfellene at data ikke kan publiseres av personvernhensyn. Enten har det ikke blitt innhentet samtykke til publisering av forskningsdata, eller så mener personvernombudet ved NSD at det ikke er mulig å publisere data, eller så er det grunnet i «regulatory conditions». I 10 av tilfellene blir det oppgitt at man kan ta kontakt med korresponderende forfatter og be om tilgang, men det gis ingen garanti for at man får se dem.

Videre var det en studie som ikke hadde samlet inn rådata, og en studie som hevder at datasettene skal publiseres i NSDs arkiv (men til dags dato ikke har gjort det). I mange av tilfellene er det vel å merke publisert både spørreskjema og intervjuguide, men uten rådata på individnivå er det umulig å gjenskape disse studiene.

Så finnes det faktisk en av 20 studier som publiserte *nesten alle* rådataene de har benyttet. Dataene består av svar på en spørreundersøkelse og her er både anonymiserte rådata på individnivå og aggregerte målinger for å svare på problemstillingene deres med. Men noe mangler for å gjøre materialet komplett: variablene for alder og kjønn er tatt helt bort, sannsynligvis av anonymiseringshensyn. I tillegg manglet spørreskjemaet som ble benyttet i studien, men dette

fant jeg til slutt etter å tatt kontakt med forfatter og blitt sendt i riktig retning via Google.

Hvorfor publisere data når man kan la være?

Så hva er årsaken til at så få studier innen dette relativt tilfeldige utvalget av artikler ikke publiserer alle sine data?

Svaret er nok i mange tilfeller forskernes mangel på kunnskap om publisering av forskningsdata og spesielt anonymisering. For forskere skal nemlig være og er nok livredde for å publisere opplysninger som kan identifisere enkeltpersoner. Spesielt når det registreres opplysninger innen særlige kategorier av personopplysninger (tidl. sensitive personopplysninger) som helseopplysninger. Verken tilliten til akademia eller fokus på åpen vitenskap er tjent med at noen ved et uhell publiserer data hvor enkeltpersoner kan gjenkjennes, ihvertfall uten deres samtykke.

Publisering av forskningsdata krever nemlig tid, nøye planlegging og ekspertise. En vanlig misoppfatning er blant annet at NSD *krever* at data slettes og anonymiseres ved prosjektslutt. Det gjør de altså ikke, selv om maler til informasjonsskriv osv. gjerne har det som standard tekst. Faktum er at det er «de registrerte», altså informantene selv som bestemmer hva deres personopplysninger skal brukes til. Som forsker kan man i informasjonsskrivet komme med et forslag til hva personopplysningene skal brukes til, informantene tar deretter stilling til om de ønsker å delta på gitte vilkår eller ikke.

Siden det i mange tilfeller ikke finnes direkte incentiver til å publisere forskningsdata, så er den enkleste løsningen å la være. Det tar for mye tid og det finnes alltid et neste prosjekt som krever oppmerksomhet.

Open Science?

I mange tilfeller er det naturligvis umulig å publisere forskningsdata grunnet kommersielle rettigheter til data, ulike lovverk for helseregistre og så videre, og jeg har på ingen måte gått dypt gjennom artiklene i utvalget for å undersøke hvorvidt det faktisk kunne la seg gjøre eller ikke. Poenget med denne bloggposten var å gi et lite innblikk i landskapet for open science innen et lite utvalg artikler publisert i ulike tidsskrift hos en stor utgiver med det jeg vil kalle en moderat open data policy. Resultatet var noe nedslående, men ikke overraskende.

Open science krever nemlig ikke bare en innsats fra forskere. Det krever enormt av institusjonene i form av støtteapparat og ekspertise. Jeg vil hevde at de aller færreste forskere har tid til å gjøre alle sider ved forskningsprosessen vidåpen for alle, noe enkelte forskningsinstitusjoner i Norge er i ferd med å forstå.

Forskningsdata