Det NFR-finansierte prosjektet «Historiske registre» (herunder «Historisk befolkningsregister») søker å utvide den nasjonale registerinfrastrukturen bakover i tid. Hovedmålet er å dekke hele perioden fra 1801 med et mest mulig fullverdig befolkningsregister som knytter sammen individer over livsløpet på tvers av kildene.

En hovedutfordring i prosjektet er transkribering av store volum av skannede håndskrevne kilder. Dette foregår i all hovedsak med maskinelt, men maskinlæringsteknikker. En hovedutfordring er å tilpasse deteksjons-teknikker til å analysere skjemastrukturer. Standard rutiner kan bukes (Detectron2, mmdetection), men det er helt avgjørende med gode tilpasninger til de gitte strukturene i våre kilder. Lesing av informasjonen gjøres delvis ved standard HTR-arkitekturer (PyLaia, mmocr, trocr), men også med rene klasse-modeller (ResNet50) for gitte typer informasjon (datoer, kodete felt etc). Det har vært en stor utfordring å etablere store treningsdatasett fra det aktuelle tekstkorpuset. Generelt har det vært helt sentralt å etablere en effektiv arbeidsflyt der maskinelle metoder og manuelt arbeid med å styrke treningsdata kan foregå mest mulig effektivt. I denne prosessen har det vært brukt ad-hoc versjoner av semi-overvåket læring.

Hovedutfordringen i bruken av maskinlæringsteknikker er vært 1) veldig høye krav til sluttproduktet og 2) stort volum. Standarden har vært å få resultater helt på linje med eller bedre enn ved manuelt arbeid. Dette krever betydelig kontekstuell tilpasning og bruk av smarte teknikker for validering. Prosjektet er et av de aller største i sitt slag, også internasjonalt. Det transkriberes trolig noe slikt som 100 millioner personforekomster fra 30 millioner bilder. Dette setter store krav til GPU-tid, men det er vel så utfordrende at heterogeniteten jevnt over også vokser i volum.

Prosjektet er et samarbeid mellom Norsk Regnesentral, Arkivverket, Statistisk Sentralbyrå, Universitetet i Tromsø, Nasjonalbiblioteket og Folkehelseinstituttet. Arbeidet med maskinlesing har i all hovedsak vært ledet av og utført av Folkehelseinstituttet.

Kontakt: Kåre Bævre