Tema 4: Formidlingen

Att ge tillgång till äldre tryckta statistikpublikationer i dagens digitala värld vid tre nordiska statistiska centralbyråer

Ragnhild Rein Bore, Statistisk sentralbyrå, Personal- og formidlingsavdelingen, rrb@ssb.no
Rolf-Allan Norrmosse, Statistiska centralbyrån, Biblioteket, rolf-allan.norrmosse@scb.se
Anna Dorthe Bracht Nielsen, Danmarks Statistik, Bibliotek og Information, abn@dst.dk

Providing access to printed statistical publications in the digital age is becoming increasingly common among statistical agencies. This paper presents the different approaches found in Norway, Sweden and Denmak. In Norway, all statistical publications will by the end of 2011 be scanned, OCR-treated and published online at www.ssb.no/histstat/publikasjoner in PDF-format through a project based on the re-allocation of internal resources. In Sweden, the series "Bidrag till Sveriges officiella statistik (BiSOS)", 1858-1910, has through external financing been digitalized by a company in Riga and made available online at www.scb.se/bisos. In Denmark, emphasis has been on registering the publication and attendant metadata in the online catalogue, www.dst.dk/katalog, in order to increase demand and knowledge of its statistical collection.

Statistisk sentralbyrån: Fra populære publikasjoner til storskala digitalisering

Fase 1: Prøveprosjekt
Prosjektet "Mer tilgjengelig historisk statistikk" hadde sin spede begynnelse i 2005 og var et initiativ fra SSBs bibliotek- og informasjonssenter. I prosjektets første fase ble alt arbeid gjort som ekstraoppgaver utenom vanlige arbeidsoppgaver. Følgende valg ble tatt:

- Publikasjonen skal vises som én PDF-fil, ikke som enkeltsider i html. Derfor viktig å produsere PDF-er i mest mulig komprimert størrelse.
- Publikasjonen skulle være i svart-hvitt, fokus er på gjengivelse av tekst og tabeller.
- Publikasjonen skal være søkbar, den OCR-behandles derfor for å gjøre teksten søkbar.
- De mest populære publikasjonene digitaliseres først.

At man i prosjektets første fase valgte å prioritere mye brukte og populære publikasjoner var for å raskt få muligheten til å ta i bruk de digitaliserte publikasjonene i analysepublikasjoner, i kurs om historisk statistikk og for å gi et bedre tilbud til brukerne som kontaktet bibliotek- og informasjonssenteret. Valget falt på:

- Folke- og boligtellingspublikasjoner fra perioden 1865-1980
- Statistiske årbøker
- Medisinalberetninger fra perioden 1856-1920
- Eldre lønnsstatistikk
- Enkelte jordbrukstellinger
- Amtmennenes femårsberetninger fra perioden 1885-1920
- Valgstatistikk

Fase 2: Storskala digitalisering av hele samlingen
I 2007 ble beslutningen tatt at hele SSBs samling av publikasjoner i serien *Norges offisielle statistikk* samt andre serier utgitt av Statistisk sentralbyrå skulle skannes og gjøres tilgjengelig på ssb.no innen 2012 slik at de kunne være en ressurs fram mot Norges grunnlovsjubileum i 2014. Hovedtyngden av digitaliseringen foregår i perioden 2009 – 2012. I alt skal ca. 17 000 publikasjoner digitaliseres. Å sette ut jobben eksternt ble vurdert, men valget falt til slutt på å gjøre jobben internt ved at en rekke ansatte i større og mindre grad fikk fristilt tid til å delta i prosjektet. I dag er det ca. 25 personer involvert i prosjektet, og i 2009 ble det brukt 10 årsverk. De mest populære publikasjonene skulle ikke lenger prioriteres, i stedet har man tatt serie for serie, og jobbet

seg bakover i tid for å kunne få nyttig erfaring med alle delene av prosjektet før det vanskeligste materialet skal digitaliseres.

Prosessen
Del 1: Pakking: Gjennomgang, slakting og pakking av papirpublikasjonene i stamarkivet
Utfordringene har inkludert: Rydding og strukturering av samlingen, velge ut eksemplarer godt egnet til slakting og skanning, komplementere samlingen ved mangler, behandling av soppskadete publikasjoner. Ansvar: Arkivet i Oslo, Seksjon for informasjon og dokumentasjon

Del 2: Skanning av publikasjonene
Utfordringene inkluderer: Riktig hardware og software, finne optimale innstillinger for ulike typer publikasjoner, grundig kvalitetskontroll m.m. Ansvar: Skannegruppa i seksjon for IT infrastruktur på Kongsvinger

Del 3: Tekstgjenkjenning: OCR-behandling og PDF-produksjon
Utfordringene inkluderer: Implementere retningslinjer for bokmerker i PDFene, valg av software, optimalisere innstillinger m.m. Ansvar: Seksjon for informasjon og dokumentasjon

Del 4: Publisering på www.ssb.no/histstat/publikasjoner
Utfordringene har inkludert: Opprette lenkesamlinger på Internett, masseoverføring av filer via FTP m.m. Ansvar: Biblioteket, Seksjon for informasjon og dokumentasjon

Del 5: Katalogisering i metadatasystem – internt og i Bibsys
Utfordringene har inkludert: Finne, anskaffe og installere metadatasystem (Tidemann), ressurser til registrering i bibliotekssystemet Bibsys. Ansvar: Biblioteket, Seksjon for informasjon og dokumentasjon

Formidling av det digitaliserte materialet
Hoveddelen av dette arbeidet vil skje etter at hele samlingen er på nett, bl.a. gjennom en mediestrategi. Strategiene hittil har inkludert:

- Presentasjonartikler på ssb.no og i fagblader
- Opptredener i ulike radioprogrammer
- Analysepublikasjoner, f.eks. På liv og død – Helsestatistikk i 150 år, samt et planlagt verk om Norge gjennom 200 år til Grunnlovsjubileet
- Kurs i historisk statistikk, 3-4 i året
- Oppføringer i Wikipedia
- Lenking fra historierelaterte nettsteder

Statistiska centralbyrån: Projekt ”Digitalisering och tillgängliggörande av bokverket Bidrag till Sveriges officiella statistik (BiSOS)”, 1851/55-1910

SCB började 2005 fundera på att digitalisera Bidrag till Sveriges officiella statistik (BiSOS), vilket är det viktigaste och mest omfattande bokverket med svensk officiell statistik från 1800-talets mitt till 1900-talets början. BiSOS – en guldgruva för studiet av Sverige och dess kontakter med omvärlden – är indelat i 23 ämnesområden/serier och fördelat på 1 495 häften eller 145 000 sidor text och tabeller i svartvit och färg. Verket togs fram under senare delen av 1800-talet under en tid med enorm teknikutveckling. Exempel på detta är järnvägar, kanaler, ångkraft, telegraf och handel. I BiSOS finns inte enbart siffror utan även beskrivning och analyser av samhällsutvecklingen.

Projektet digitalisering började med att två konsultföretag dels digitaliserade en mindre del av BiSOS, dels gjorde en utredning rörande digitalisering av hela BISOS med arbetsflöde, tidsåtgång,

kostnader, etc. 2006 ansökte SCB om infrastrukturellt stöd hos Riksbankens Jubileumsfond (RJ) för digitalisering av BiSOS. RJ beviljade SCB 4 miljoner kr som skulle bekosta digitalisering, dvs. bildfångst, teckenigenkänning och framställning av presentationsfärdig pdf-publikation. SCB skulle bekosta upphandling, preparering av materialet, kvalitetssäkring och publicering på SCB:s webbplats. Efter genomförd upphandling tecknades avtal i juli 2007 med Logica. Som underleverantör använde Logica Infodisk media, Riga, Lettland.

Efter omfattande tester och granskning av leverantörens arbetsmetoder och rutiner och fastställande av standarder som skulle användas kunde SCB:s arbete med kvalitetssäkring minimeras. Innehållsförteckningar är viktiga och därför ställde SCB kravet på korrekthet av OCR-tolkning på ord- och teckennivå till 99,9 %. I slutet av 2007 godkände SCB de första serierna.

Några exempel på standarder:
- Olof Dahlins ordbok, 1855, har använts som ordlista vid teckenigenkänning
- Särskilda beskrivningsblad har skapats för att kommunicera med leverantören
- En särskild inledningssida har införts för att beskriva föregångare, efterföljare och översiktspublikation samt med digitaliseringsinformation och identifikatorn urn:nbn
- Dokumentegenskaper har beskrivits med metadata

Projektet strävade att bevara det goda i det tryckta materialet och samtidigt åstadkomma en användarvänlig digital version. Resultatet blev mervärden eller förbättringar jämfört med det tryckta originalet. Nedan presenteras några förbättringar i digitala BiSOS jämfört med den tryckta förlagan:
- Innehållsförteckningar har skapats om sådan saknas
- Innehållsförteckning finns alltid i början av häftet
- Klickbara innehållsförteckningar
- Bokmärken till lägsta nivå
- Liggande tabeller och text vrids till stående

I maj 2009 var BiSOS enligt tidsplan färdigt. Kostnaden uppgick till 1,2 milj. kr. De återstående medlen på 2,8 milj. kr har RJ godkänt att SCB får använda för digitalisering av material från 1811-1857 och 1911-2001. Omfattningen är cirka 290 000 sidor.

SCB:s digitalisering finns på www.scb.se under Hitta statistik > Historisk statistik.

Danmarks Statistik: Projekt "Større synlighed af samlingen, herunder magasiner"

Danmarks Statistik har ikke som SSB og SCB valgt at synliggøre de ældste udgivelser gennem digitalisering. Det eneste, der indtil videre fuldtekstscannes, er Statistisk Årbog, som gøres tilgængelig på www.dst.dk/aarbogsarkiv og via Det Kongelige Biblioteks portal for den danske digitaliserede kulturarv kaldet Kulturperler: www.kb.dk/da/materialer/kulturarv. Dog bestræber vi os på at fuldtekstscanne de eksemplarer, vi i løbet af vores projekt finder ud af, vi kun har ét eksemplar af.

Vi har i stedet valgt at synliggøre den ældre samling helt op til 1995 ved at oprette bibliografiske poster i vores online katalog. Publikationer anskaffet før 1995 er nemlig kun registreret i kortkatalog. Ved at få nykatalogiseret bøgerne i bibliotekssystemet bliver de søgbare i vores egen Opac, dst.dk/katalog, og fordi posterne eksporteres til den danske fælleskatalog DanBib, kan den professionelle bruger fremsøge dem via Netpunkt, mens den almindelige slutbruger kan søge i bibliotek.dk. På den måde håber vi at skabe større efterspørgsel og bedre udnyttelse af vores samling. Brugerne skal gøres bekendt med, og gerne overraskes over, hvor mange gamle og interessante publikationer, der er i vores samling. Løsningen er valgt i erkendelse af, at

publikationer, der ikke kan findes i en online katalog, for mange brugere simpelthen er ikke-eksisterende. Hvis vi på et senere tidspunkt får ressourcer til at fuldtekstscanne, vil vi udover at kunne gøre materialet tilgængeligt via links på vores hjemmeside, også have den registrering, som gør det søgbart via landets biblioteker.

Selvom det ikke er et digitaliseringsprojekt, vil vi gerne give brugerne fornemmelsen af de gamle bøger og en god mulighed for at relevansvurdere materialet. Derfor scannes indholdsfortegnelserne til pdf-filer og tilknyttes katalogposterne. Da mange af fortegnelserne kan være meget lange, OCR-scannes de, så teksten i filen bliver søgbar. Yderligere hjælp til relevansvurdering får man i den indholdsbeskrivende note, som vi udarbejder for hver titel, men den er i lige så høj grad tilknyttet for at sikre posterne gode, relevante emneord og optimere søgningen. Vores eget emneords- og klassifikationssystem til tildeling af kontrollerede emneord er nemlig ikke særlig detaljeret.

Projektet startede i 2008, og med én deltidsansat, der udelukkende har arbejdet med katalogisering og tre andre, der sideløbende med bibliotekets øvrige arbejde har deltaget, er vi nu færdige med alle folianter (26 titler), monografier (463 titler) og 4 periodika-titler (1263 eksemplarer). Som det første gik vi i gang med monografierne, da det var nyttigt både for brugerne og os selv at få synliggjort de mange enkeltstående titler bl.a. i serien "Statistiske Undersøgelser". Det er også en lettelse nu at have tydet alle de gotiske bogstaver i de alenlange titler i folianterne, og nu endelig have et overblik over, hvad de indeholder. Det er dér, den ældste statistik er offentliggjort, og der er ofte bud efter at vide, hvor langt tilbage man kan komme i tid. De store serier i "Statistisk Tabelværk" fx "Ægteskaber, Fødte og Døde", Folketællingerne og Erhvervstællingerne er også katalogiserede og eksemplarerne stregkodet.

At eksemplarerne er stregkodet betyder, at vi nu har overblik over bestanden af de gamle publikationer og kan se hvor mange eksemplarer vi har af de enkelte titler. Det har ikke været muligt hidtil, da vi kun har haft en uopdateret note på bagsiden af kartotekskortene. Det viste sig, at vi havde så mange eksemplarer, at vores regel om ikke at udlåne bøger fra før år 1900 kunne ophæves. Der er da også særlig interesse for disse de ældste bøger, og de udlånes til både biblioteker og direkte til privatpersoner. Titler, vi kun har ét eksemplar af, udlånes ikke, men som tidligere nævnt bestræber vi os på at fuldtekstscanne dem, så de er tilgængelige online.

Projektet har således allerede givet resultater. Vi har da også aktivt gjort opmærksom på, at vores katalog nu indeholder alle disse nye poster og de forbedrede muligheder for at finde historisk statistik ved en række artikler om projektet og om historisk, statistisk informationssøgning. Artiklerne er publiceret i fagbladet Referencen, som er rettet mod bibliotekarer, der ude på landets biblioteker hjælper brugerne med informationssøgning.

Afslutning

De nordiske statistikbureauer har store mængder af data til forståelse af samfundets udvikling gennem en næsten 200-årig periode. Internettet og teknikken til digitalisering giver mulighed for at tilgængeliggøre og formidle disse data til gavn for brugerne. Brugerne efterspørger de gamle statistikker, men deres forventning er samtidig at alt findes på Internettet. Uanset hvor mange ressourcer man har til rådighed, er det muligt at lave tiltag der forbedrer de gamle bøgers tilgængelighed; det har SCB's, SSB's og DST's tre projekter vist.