Tema: Formidling
Sanne Vestergaard Nielsen, svn@dst.dk
Danmarks Statistik

Begreber – et projekt om forklaring af statistik

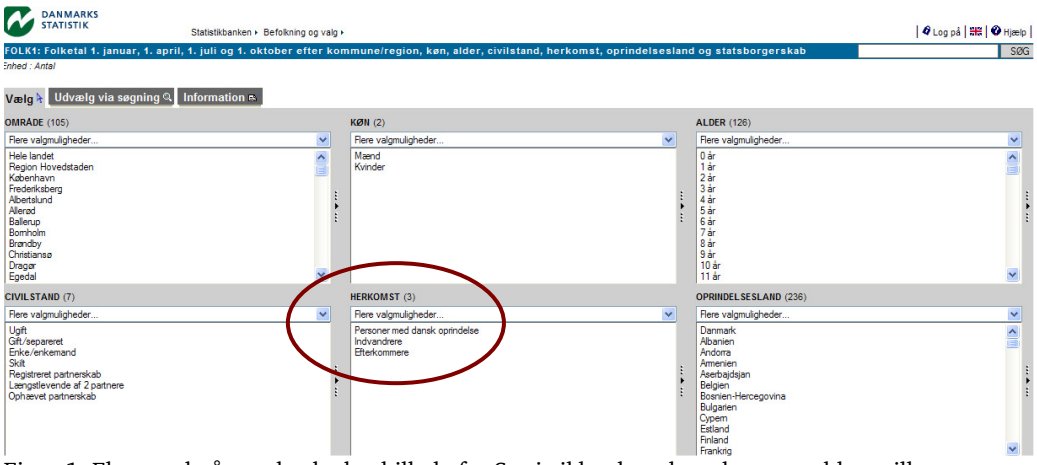
Abstract Danmarks Statistiks databanker har i de seneste år arbejdet med en begrebsdatabase, der med udgangspunkt i Statistikbanken, vil forklare begreber i statistikken for brugerne. Resultatet af arbejdet bliver en A-Z ordbogsløsning på Danmarks Statistiks hjemmeside, samt en del knyttet til Statistikbanken. Præsentationen vil tage udgangspunkt i de problematikker og overvejelser vi har måtte gøre os undervejs, styrker og svagheder på det endelige produkt, samt fremtidige udvidelsesmuligheder.

Baggrund Som et led i at vurdere om Danmarks Statistik lever op til brugernes behov om let adgang til statistik og for at få input til, hvad vi bør gøre bedre, foretager vi hvert år en brugertilfredsundersøgelse af Statistikbanken. [1]

I mange år har der i disse brugerundersøgelser været et ønske om bedre dokumentation. I brugerundersøgelsen fra 2007 var 20,5 procent af brugerne enten utilfredse eller meget utilfredse med dokumentation og forklaringer til tabellerne[2]. Desuden ytrede brugerne et ønske om forklaringer på de enkelte værdier i en tabel. Fx Hovedstadsregionen – hvad dækker det over etc.

Projektet opstod således ud fra vores systematiske målinger af brugertilfredshed og af den daglige feedback. Det blev formuleret et behov, som det blev besluttet at løse. Derfor var det i høj grad brugernes behov, ikke de bagved liggende systemer, der blev afgørende for projektets form.

Begyndelsen Som vi analyserede brugernes kritik, manglede brugerne en form for dokumentation af de enkelte ord og begreber i brugen af Statistikbanken. Konkret i udvælgelsesbillet i Statistikbanken havde brugerne ikke mulighed for at foretage kvalificerede valg. Fx hvad er en "Indvandrer", hvad er en "Efterkommer" og hvori ligger forskellen.



Figur 1. Eksempel på et udvælgelsesbillede fra Statistikbanken, hvor brugeren bl.a. stilles over for valget mellem "Indvandrere" og "Efterkommere".

I Danmarks Statistiks forskellige trykte publikationer findes ofte stikord og begrebsforklaringer, der, som en del af dokumentationen af tællingen, forklarer

[1] Seneste brugerundersøgelse i denne form er dog 2007, grundet en omlægning.
[2] 20,5% af de brugere der har trukket mere end et tal fra Statistikbanken, og 18,4% af de brugere der kun har trukket et tal.

begreberne. Men oplysninger, der skal findes i bog eller et andet sted på internettet, hjælper ikke statistikbankens brugere. Udgangspunktet var således at tage elementer fra allerede eksisterende kilder og skabe en sammenhæng til Statistikbanken.

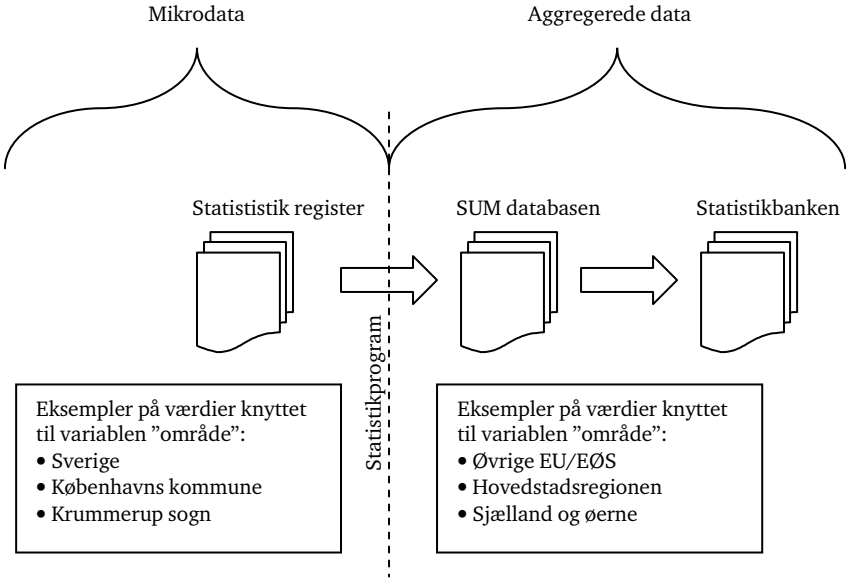
Typer af forklaringer

Efter at have analyseret udvælgelsesbilledet viste det sig, at der var tale om to primære typer af forklaringer der ville tilføje værdi for brugeren.
1) Overordnede forklaringer, der knytter sig til en udfaldsvariabel.
2) Mere konkrete forklaringer, der knytter sig til en værdi i Statistikbankens udvælgelsesbillede.

I Figur 1 ville en forklaring knyttet til "Herkomst" være af den overordnede type. En mere konkret forklaring ville knytte sig til de enkelte værdier under "Herkomst", dvs. "Indvandrere" og "Efterkommere".

De overordnede forklaringer er relevante hvor der f.eks. er anvendt nomenklaturer. De enkelte værdier i nomenklaturen kan være selvforklarende, men en henvisning til hvilken nomenklatur der anvendes, og hvilken sammenhænge den anvendes i, vil give brugeren en bedre tilgang til statistikken. Denne type information findes i høj udstrækning allerede, men som en del af dst.dk/dokumentation, ikke som en del af statistikbanken.

De konkrete forklaringer, knyttet til værdier i Statistikbankens udvælgelsesbillede, har været den største opgave. Som type er det en meget konkret forklaring, der skal hjælpe brugeren til at udvælge de korrekte værdier. Som det tidligere eksempel med indvandrere og efterkommere.



Figur 2. Eksempel på mikrodata og aggregerede data i Statistikbanken.

Målgruppe

Halvdelen af Statistikbankens brugere er studerende, skoleelever eller på anden vis tilknyttet en uddannelsesinstitution. Udover det er 15% privatpersoner. Kun 6% er forskere, mens i alt 30% af brugerne er fra virksomheder, statsinstitutioner eller kommuner.[3] Det er klart, at disse tal har haft stor betydning for overvejelserne omkring begrebsforklaringerne. For at kunne ramme så stor en del af brugerne er målet at bruge et "almindeligt sprog" og så vidt muligt undgå de tekniske formuleringer.

[3] Tal pr 1. juni 2010

En anden pointe er, at selv ord som vi betragter som selvforklarende er netop ikke selvforklarende for en stor del af vores målgruppe. Særligt ved aggregeringer opstår der et gab mellem brugernes informationsbehov og det produkt vi kan tilbyde dem. Behovene er ofte konkrete, og kan ikke besvares i den information der ligger i udvælgelsesbilledet. ”Jeg skal bruge information på Frederiksberg Kommune. Er den inkluderet i Hovedstadsregionen?”

De eksisterende kilder

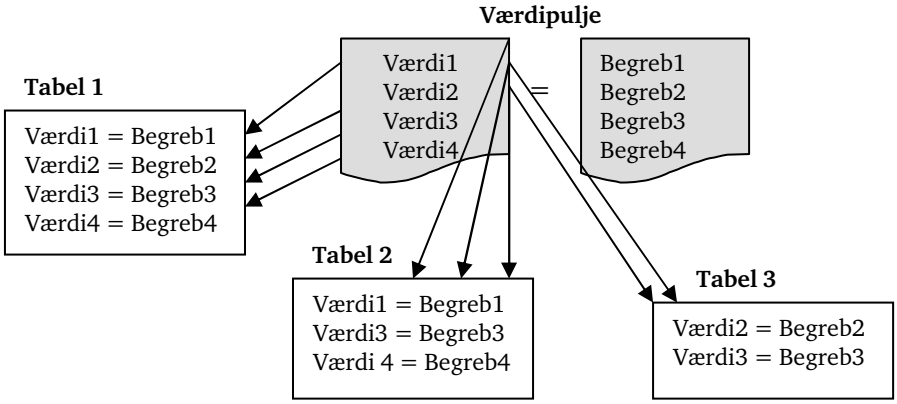
Projektets fokus er formidlingen. Altså at kunne præsenterer de informationer vi allerede har, på det sted hvor brugerne efterspørger det. Derfor har de eksisterende kilder til dokumentation at statistikkerne været i centrum hele vejen igennem.

De vigtigste kilder har været de der retter sig mod målgruppen, af hensyn til sproget og forståelsen, men også Times, Danmarks Statistiks dokumentationssystem på mikro-niveau, har været anvendt.

Datamodel

Der har været flere bud på en datamodel undervej i projektet. Oprindelig var ideen, at hver værdi i en tabel skulle have sin egen begrebsforklaring. Det der talte i mod denne løsning var, at i Statistikbanken er metadata fælles. Dvs. at den samme værdi optræder i flere tabeller. Det store spørgsmål er så, om en værdi altid betyder det samme, alle de steder den optræder i Statistikbanken.

Løsningen blev at datamodellen blev knyttet til de metadata som allerede eksisterer i Statistikbanken. Dvs. at en begrebsforklaring hænger sammen med en værdi, som den ligger i en værdipulje.



Figur 3. Ved at knytte et begreb til værdien i datamodellen, bliver den samme begrebstekst anvendt flere steder.

En af fordelene ved dette er at redigeringen af begrebsforklaringerne hænger sammen med redigeringen af værditekster. Således har de interne brugere i Danmarks Statistik allerede en forståelse for, hvordan tingene er forbundet.

Ulempen er, at der ikke ubesværet kan laves undtagelser. En værdi kan ikke have forskellige begrebsforklaringer i forskellige tabeller. Spørgsmålet er om et ord altid betyder det samme, eller om betydningen kan varierer fra en sammenhæng til en anden. Vi ved, at ords betydningen ændrer sig. Det sker over tid, i ændringer i aggregeringer og i ændringer i verdensbilleder.

Dele af den metadata der ligger i Statistikbanken har ligget der siden 2002. Den er anvendt på mere end 2500 offentlige tabeller. Det er klart, at der vil være tilfælde, hvor det samme ords betydning i to forskellige tabeller har ændret sig over tid. Det er prisen man betaler for at tilføje data til noget der har haft sit eget liv i omkring 10 år. Forhåbentlig er problemet ikke ret stort, men omfanget vil først blive klart, når begreberne bliver lagt på Statistikbankens grænseflade.

To anvendelsesmuligheder for data

Undervej i projektet gik det op for os, at der kunne være flere anvendelsesmuligheder end i Statistikbanken. Dokumentationen kunne måske genanvendes på dst.dk som en type ordbog? Tanken var at kunne genbruge data med en relativ stor grad af automatik og genbrug.

Den kom til at hedde "Hvad betyder… og er i dag i funktion på dst.dk/hvadbetyder.

Under arbejdet med at udvikle grænsefladen til "Hvad betyder…" opstod der en række problemer.

Eksempelvis tager begrebsdatabasen udgangspunkt i de ord og begreber der er præsenteret i Statistikbanken, men vi forudså allerede på et tidligt tidspunkt behovet for at kunne præsenterer ord og begreber der ikke findes i Statistikbanken. Ord som "Gini-koefficient" eller "Kvartiler". I praksis betyder det en tredje type af begrebsforklaringer;
3) Begrebsforklaringer til ord der ikke findes i Statistikbanken

En anden problemstilling var, at ved automatisk at trække materialet tiltænkt Statistikbanken til et andet medie, kan kontekstændringen have større konsekvenser end først tiltænkt. Et ord som "Kapacitet" vil have en given betydning i Statistikbanken set i forhold til den tabel ordet forekommer i, men i "Hvad betyder…" står ordet alene, og den forklaring der kunne passe i Statistikbanken ikke er fyldestgørende i "Hvad betyder…"

"Hvad betyder…" har en alfabetiseret del der er afhængig af at opslagsordet er konstrueret med den betydende del først, men Statistikbanken indeholder en del værdier, hvor dette ikke er tilfældet. "Andre forbrugsvarer" og "Antal bedrifter" bare for at nævne et par stykker.

Figur 4. Illustration af problemet med at alfabetiserer ting, der er taget ud af en anden kontekst.

Hvor langt er vi kommet?

Visionen var at Statistikbankens brugere ville få mulighed for at få præsenteret den data de har brug for, hvor de mangler den. 1. november i år bliver begrebsdatabasen lagt på Statistikbanken, så der når vi et godt stykke videre mod målet.

Det betyder ikke, at arbejdet er overstået. Tværtimod. Selvom kilder til alle statistikområder har været gransket, så er der et stykke til, at alle tabeller er dækket så godt som man kunne ønske. Dertil komme det arbejde, der ligger ud over det konkrete produkt. Alt det arbejde der handler de spørgsmål dette projekt har genereret, snarere end de projektet har besvaret.