SCB:s erfarenheter av digitalisering av Bidrag till Sveriges officiella statistik (BiSOS)

Rolf-Allan Norrmosse

Paper presenterat vid konferensen

Mötesplats inför framtiden
ARBETSLIV·UTBILDNING·FORSKNING

14–15 oktober 2009 i Borås

BORÅS STAD
Stadsbiblioteket

HÖGSKOLAN I BORÅS
Institutionen Bibliotekshögskolan

Statistiska centralbyrån
Statistics Sweden

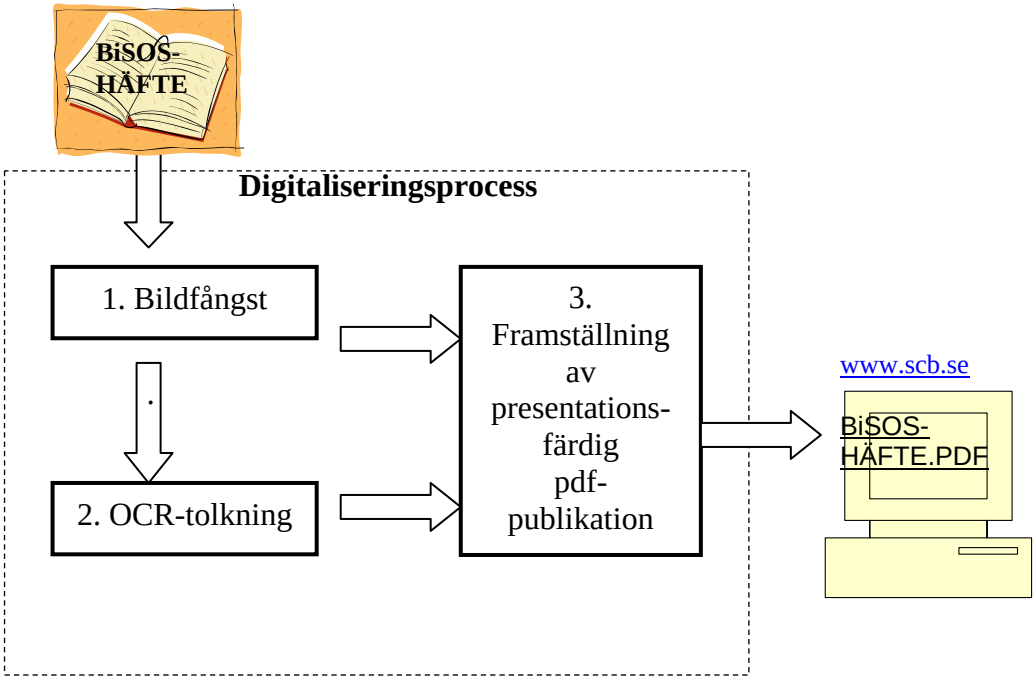
SCB:s erfarenheter av digitalisering av Bidrag till Sveriges officiella statistik (BiSOS)

På Mötesplats inför framtiden 2007 berättades om vikten av arbetet med att digitalisera BiSOS.

- BiSOS är det viktigaste och mest omfattande bokverket med svensk officiell statistik från 1800-talets mitt till 1900-talets början.
- BiSOS är en guldgruva för studiet av Sverige och dess kontakter med omvärlden. BiSOS tillhör forskningens infrastruktur och är en del av Sveriges kulturarv.

I januari 2006 ansökte SCB om 4 miljoner kr hos Riksbankens Jubileumsfond (RJ) för digitalisering av BiSOS. Medlen beviljades i oktober 2006 för upphandling av tjänsten enligt *digitaliseringsprocessen* se bilden nedan.

Digitaliseringsprocessen består av följande:
1. Bildfångst
2. OCR-tolkning
3. Framställning av presentationsfärdig pdf-publikation



Att digitalisera BiSOS var komplext och innebar allehanda utmaningar både för SCB:s bibliotek och för leverantören SIA Infodisk media, Riga, Lettland, som är underleverantör till Logica.

SCB:s leveransplan var på 18 månader med sista serien klar i april 2009. Den 28 maj 2009 hade vi avslutningsmöte med Logica och Infodisk media.

BiSOS är nu tillgängligt på www.scb.se under Hitta statistik – Historisk statistik. – BiSOS eller direkt via www.scb.se/BiSOS. BiSOS finns i Libris på häftesnivå, både digital och tryckt version.

BiSOS har länge inspirerat både forskare och allmänhet till studier inom de mest skilda företeelser i samhället. Tillgängliggörande över internet har redan ökat användningen av verket. En del tidigare mindre använda serier har resulterat i rapporter vid universiteten. Studenter vid universitet har kontaktat projektet för att få kännedom om när vissa serier publiceras digitalt. Intresset för BiSOS har också märkts på Bok- och biblioteksmässan.

Även utanför Sverige har användare upptäckt digitala BiSOS, bl.a. i Australien och Kanada. Våra nordiska systerbibliotek och andra bibliotek har rapporterat nyttan av att BiSOS nu finns tillgängligt.

Utveckling och förnyelse vid digitalisering.

När BiSOS togs fram under 1800-talet försökte man genom mallar få enhetlighet i framställningen och det märks att man tänkte på användaren. Projektet strävade att bevara det goda i det trycka materialet och samtidigt åstadkomma en användarvänlig digital version. Detta berördes redan i upphandlingsunderlaget där vi ställde krav på användande av standarder och utvecklade tekniker för att stärka användarvänligheten. Resultatet blev mervärden eller förbättringar jämfört med det tryckta originalet.

Här ges några exempel på standarder och utveckling:

- Granskat leverantörens arbetsmetoder och rutiner för att minimera SCB:s arbete med kvalitetssäkring.
- Olof Dahlins ordbok 1855. Använd som ordlista vid teckenigenkänning (OCR-tolkning).
- Utnyttjat en den unik identifikatorn URN:NBN för häften och bilder.
- Bevarat BiSOS paginering i PDF:ens sätt att paginera.
- Skapat beskrivningsblad för att kommunicera med leverantören av tjänsten.
- I den digitala versionen infört en särskild inledningssida för att beskriva föregångare, efterföljare och översiktspublikationer, samt digitaliseringsinformation och URN:NBN.
- Dokumentegenskaper med metadata.
- Skapat innehållsförteckningar som saknas i 477 av de 1495 häftena.
- Infört start settings för att presentera det digitala häftet på ett optimalt sätt för användaren.

Förbättringar i digitala BiSOS jämfört med den tryckta förlagan:

- Skapad innehållsförteckning om den saknas
- Innehållsförteckning alltid i början av ett häfte
- Klickbara innehållsförteckningar
- Länk till annan innehållsförteckning eller tabellförteckning i innehållsförteckningen
- Bokmärken till lägsta nivå
- Liggande tabeller eller text vrides till stående
- Serie- och/eller volymuppgift visas alltid på titelsidan även om den saknas i den tryckta förlagan
- Inledningssida – vänstersida till titelsidan - med föregångare, efterföljare och översiktspublikation tillika med urn och digitaliseringsinformation

Kraven på OCR-tolkning fastställdes till följande krav på korrekthet på tecken- och ordnivå:

- Svenska innehållsförteckningar: 99,9 %
- Franska innehållsförteckningar: 99,5 %
- Annan text: 85 %. (Denna korrekthet bestämdes efter studiebesök vid Nationalbiblioteket i Oslo).

Det var inte möjligt att inom projektets ekonomiska ramar OCR-tolka tabellerna med tillfredsställande kvalité. Det var bättre att spara alla bildfiler i formatet tiff, vilket är viktigt av flera skäl. Formatet är arkivgodkänt av Riksarkivet. Vi har tillgång till originalen och kan, när tekniken medger, göra en bra OCR-tolkning av tabeller. Detta kommer vi att få nytta av i vårt kommande projekt: ”Förstudie till Verktyg för analys och textigenkänning av tabeller”. (Se nedan).

Ett särskilt problem med text från tiden för BiSOS publicering är att stavningssättet var ett annat än det som används idag. Språkdata vid Göteborgs universitet har Olof Dahlins ordbok från 1855 i digital version. SCB har fått lov att använda denna i BiSOS digitalisering. För att få den kvalité vi eftersträvar på OCR-tolkningen jämför man alla ord med en ordlista. Om ordet inte finns i ordlistan visas det för OCR-operatören som då manuellt godkänner alternativt korrigerar ordet. Detta har också inneburit att arbetet med att korrigera OCR-tolkningen har tagit mer tid i början av varje serie än senare. På detta sätt har vi fått betydligt högre kvalité än det som krävdes i upphandlingsunderlaget. Dahlins ordbok blir kompletterad med nya ord och det är denna ”nya” Dahlins ordbok som vi skickat tillbaka till Språkdata för framtida forskning.

Eftersom BiSOS publicerades under mer än 50 år finns det självklart en hel del variationer i hur redovisningen har skett inom de olika ämnesområdena. Leverantören behöver få hjälp att arbeta på ett enhetligt och systematiskt sätt. För kommunikation mellan SCB och leverantören har för varje häfte tagits fram ett beskrivningsblad. Detta blad innehåller uppgifter om:

- Den unika identifikatorn URN:NBN, filnamn och annan metadata, som skall finnas i pdf-publikationens dokumentegenskaper
- Metadata, som ska finnas i vissa bokmärken, t.ex. titel.
- Bibliografiska uppgifter eller metadata som skall användas vid registrering i databasen Libris.
- Noggrann beskrivning av hur häftet är uppbyggt, paginering, antal sidor och utvik i svartvit, gråskala och färg
- Speciella avvikelser och skador i ett häfte
- Namn på filer för skapade innehållsförteckningar och inledningsblad

Kort sagt är beskrivningsbladet det enda leverantören behöver för att kunna utföra sitt arbete – där står allt! Finns oklara detaljer skickas en förfrågan till SCB. Leverantörens kvalité i arbetet liksom SCB:s har förstärkts.

Vår leverantör Infodisk media, Riga, har utmärkta kunskaper i bildfångst, OCR-tolkning och produktion av publiceringsfärdig pdf-publikation, men är också en lärande institution. Detta har bl. a. märkts vid våra möten i Riga. Efter diskussioner har vi alltid kommit fram till en optimal lösning.

Med tanke på att det var en EU-upphandling gjordes extra tydliga skrivningar i upphandlingsunderlaget om språk, rutiner, ansvarsfrågor, försäkringar, transporter, hantering av materialet, m.m.

Resultat

Projektet BiSOS digitalisering har haft som mål att informationen i ett BiSOS-häfte skall vara fullständig. Saknas en sida eller är den skadad och inte finns hos SCB har ett exemplar lånats in från annat bibliotek för digitalisering. Den digitala versionen av BiSOS-häftet har på så sätt blivit det fullständiga exemplaret av häftet.

Resultatet av projektet är att Bokverket BiSOS finns att läsa på SCB:s webbplats, www.scb.se/BiSOS. Arbetet pågår med att registrera alla BiSOS häften i forskningsbibliotekens databas Libris. Genom detta blir uppgifter om verket BiSOS tillgängliga inte enbart vid alla offentliga bibliotek i Sverige utan också vid bibliotek utanför Sverige. Det uppmärksammades tidigt från Kanada och Australien att BiSOS serie befolkning fanns tillgänglig på internet.

Ett annat resultat är att bildfiler i format tiff är av såpass hög kvalitet att vi får nytta av det i vårt kommande projekt ”Förstudie till Verktyg för analys och textigenkänning av tabeller” (Se nedan).

Av 4 miljoner kr förbrukades drygt 1,2 miljoner kr. BiSOS återstående medel uppgår alltså till 2,8 miljoner kr. En förklaring till detta är att produktionen är helt förlagd till Riga, Lettland. I digitaliseringsprocessen är särskilt OCR-tolkning och skapande av en publiceringsfärdig pdf personalintensivt. Om BiSOS digitalisering hade utförts i Sverige hade alla medel använts beroende på det högre löneläget i Sverige.

Vad händer framöver - Användning av BiSOS återstående medel

SCB har skrivit till Riksbankens Jubileumsfond och lämnat förslag på användning av BISOS återstående medel. RJ har beslutat att SCB får disponera återstående medel under 2009 och 2010 med slutrapportering 24/3 2011. Medlen ska användas till:

Verktyg för analys och OCR-tolkning av tabeller

Att kunna överföra tabelldata från bilder (tiff-filer) eller pdf-publikationer till databaser eller kalkylark efterfrågas starkt av våra kunder. Syftet med förstudien nu är att undersöka, om det finns något verktyg tillgängligt på marknaden för analys och teckenigenkänning av tabeller som kan anpassas till en statistikproducents behov. Om sådana verktyg inte finns ska förstudien beräkna kostnaden för att SCB utvecklar verktyget. I förstudien skall från det material som SCB har digitaliserat tas fram ett antal typfall på hur tabeller har varit uppbyggda sedan 1811. Med verktyget skall tabeller kunna teckentolkas med inga eller ett fåtal fel. Kontrollräkning skall kunna göras.

Statistiska centralbyrån
Statistics Sweden

Fortsatt digitalisering av tryckta statistiska publikationer

Fortsatt digitalisering av äldre statistik både från tiden före och tiden efter BiSOS. Bland annat Statistisk årsbok för Sverige och Folk- och bostadsräkningarna. Dessutom användbar statistik mellan 1900 och fram till 1962, då centraliseringen av statistiken till SCB påbörjades.

Rolf-Allan Norrmosse, SCB
rolf-allan.norrmosse@scb.se

Statistiska centralbyrån
Statistics Sweden

Fortsatt digitalisering av tryckta statistiska publikationer.

SCB vill använda större delen av återstående medel till fortsatt digitalisering av tryckta statistiska publikationer. Precis som för BISOS tillhör denna statistik forskningens infrastruktur och även Sveriges kulturarv. Genom digitalisering kan originalpublikationer skyddas från slitage och bevaras. Att den digitala versionen blir fritt tillgänglig över internet är till nytta inte enbart för forskningen utan även för den enskilde medborgaren.

Tablå

Kostnaden för digitalisering av tryckta publikationer är beräknad med den mall som används för BISOS digitalisering. Logica kommer att ha samma prissättning som för BISOS. Kostnaden är beräknad till 2 miljoner kr.
Blå rader betyder att serien är publicerad eller klar att publicera på SCB:s webbplats.

	Bokserie eller tidskrift	**Period**
1	Föregångare till BISOS. 29 000 sidor. 8 av BISOS 23 serier har föregångare. Den mest omfattande serien är H Kungl. Maj:ts befallningshafvandes femårsberättelser. Många utvik.	1811-1867
2	Statistisk årsbok för Sverige. 45 000 sidor	1914-1998
3	Folk- och bostadsräkningen. 2 000 sidor.	1914-1998
4	Valstatistik. 21 000 sidor.	1911-2001
5	Befolkning. I huvudsak. 50 000 sidor	1911-2001
6	Index till Statistiska meddelanden, 1963-2006. A4. 4 000 sidor.	1963-2006
7	[Statistiska centralbyråns stencilerade publikationer]. A4. Stencilerat. 14 000 sidor	1952-1963
8	Statistiska meddelanden. Ser A. Tillfälliga statistiska undersökningar. 4 000 sidor	1912-1953
9	Statistiska meddelanden. Ser. C. Månadsstatistik över handeln. 12 000 sidor	1912-1953
10	Statistiska meddelanden. Ser. D. Järnvägsstatistiska meddelanden. 7 000 sidor	1913-1954
11	Statistiska meddelanden. Ser. E. Uppgifter om bankerna. 11 000 sidor	1912-1953
12	Sociala meddelanden (Statistiska meddelanden. Ser.F). 15 000 sidor	1912-1967
13	Statistisk tidskrift, inkl Sveriges officiella statistik i sammandrag 1870-1913. Sammandraget är föregångare till Statistisk årsbok för Sverige. 15 000 sidor	1860-1912
14	Arbetsstatistik. Serier: A-F, H, L. Utgifvna af K. Kommerskollegii afdelning för arbetsstatistik. 4 000 sidor	1899-1910
15	Meddelanden från K. Kommerskollegii afdelning för arbetsstatistik. 9 000 sidor	1904-1911
16	Ekonomisk översikt / utarbetad inom Kommerskollegium. 15 000 sidor	1899-1910
17	Kommersiella meddelanden / utgiven av Kungl. kommerskollegium. 37 000 sidor	1914-1962
18	Återstående monografier i serien Historisk statistik för Sverige.	