Regesta Norvegica, fra tradisjonell utgave til elektronisk versjon

Innlegg holdt på seminaret 'Modernisering av tradisjonell kildeutgivelse',
XX Nordiske arkivdager, 08.08.03

Mette Gismerøy Ekker

([ill. + nr.] viser til forklarende illustrasjon i tilhørende powerpointpresentasjon)

1. Innledning

Ved Kjeldeskriftavdelingen ved Riksarkivet i Oslo holder vi på med vår første elektroniske publikasjon; vi lager en digital versjon av verket Regesta Norvegica.

Dette er et samarbeidsprosjekt mellom Riksarkivet ved Kjeldeskriftavdelingen, og Christian Emil Ore ved Enhet for Digital Dokumentasjon ved HF, Universitetet i Oslo. Denne enheten er Dokumentasjonsprosjektets etterorganisasjon, også kalt DOK-gruppen, som er den betegnelsen jeg kommer til å bruke i denne presentasjonen.

2. Presentasjon av det trykte verket Regesta Norvegica

Regesta Norvegica er en av seriene som utgis ved Kjeldeskriftavdelingen. Kort sagt kan vi si at den er en samling av regester eller sammendrag av middelalderdokumenter som har betydning for norsk historie. Denne samlingen er kronologisk ordnet.

Det er til nå kommet ut 7 bind i serien. De dekker perioden 822-1390, og er utgitt i tidsrommet 1978 til 1997.
Høsten 2003 skal 8. bind i serien komme, og da er verket ført fram til 1404.
Bind 1-7 inneholder nesten 9.000 regester, og med bind 8 er vi oppe i godt 10.000 regester.

2.1 Tidsperspektiv

Prosjektet digital Regesta startet opp høsten 2002, og i løpet av våren 2003 lå de 3 første bindene inne i en prøvebase. Bind 4 og 5 er klare til å legges ut i basen, og bind 6 er under arbeid. Hvis vi fortsetter i samme tempo, vil alle 8 bind med sine vel 10.000 regester være inne i prøvebasen i løpet av året.

3. Regesteksempel 1

Vi skal kort se på hva en typisk regest kan inneholde.
[ill. 5]
Dette er regest nr. 467 i bind 3 av Regesta Norvegica.

- Regesta sier i sin målsetting at verket skal gi korte sammendrag (dvs. regester) av innholdet i hvert dokument.

Her har vi et gavebrev fra kong Håkon Magnusson. Det gjelder donasjon av jord til kirken. Kongen gir en gård til kannikene i Nidaros for at de skal holde sjelemesse, såkalt årtidhold, for kongen og dronningen.

- Videre skal Regesta gi opplysninger om datering og utstedelsessted, - dette finner vi i regestens overskrift.

- Verket skal opplyse om oppbevaringssted for originalen (f.eks. arkivsignatur), og trykkested (som i dette tilfellet er Diplomatarium Norvegicum). Denne informasjonen finner vi i petitavsnittet som også kan gi andre opplysninger avhengig av type original.

I tillegg ser vi at hver regest er utstyrt med et nummer. Kombinasjonen av bindnummer og regestnummer er en entydig identifikasjon av regesten.

4. Bakgrunnen for prosjektet

Det er altså denne typen materiale vi nå er i ferd med å gjøre søkbart i en database. Og hvorfor gjør vi det?
Fra Riksarkivets synspunkt er hovedgrunnen å få utvidet søkemulighet i materialet. I en Regestabase kan vi for det første søke på tvers av bindene. De fleste bindene dekker ganske korte perioder på 15-20 år, og det er derfor ikke uvanlig at en person eller sak forekommer i mer enn ett bind. For det andre vil det å utføre fritekstsøk i en relasjonsdatabase åpne for flere søkemuligheter enn det de trykte binds registre gjør.
For vår samarbeidspartner, Christian-Emil Ore og DOK-gruppen, er en Regestabase en naturlig utvidelse av Dokumentasjonsprosjektets middelalderbaser. Disse basene omfatter bl.a. Diplomatarium Norvegicum som inneholder mye av det materialet Regesta er basert på. En Regestabase kan være en innfallsport til Diplomatariebasen for brukere som ikke behersker originalens språkform.

5. Gangen i arbeidet

[ill. 7]
Vi skal se litt på arbeidsgangen i prosjektet.

Vi definerer det stadiet vi er på nå som prosjektets fase 1. Vi har valgt å gjøre ting relativt enkelt slik at vi raskest mulig kan få alle bind inn i prøvebasen.

5.1 Konvertere eller skanne

Vi jobber med ett bind av gangen, og begynner med å opprette en rentekstfil av bindet. Hvordan dette gjøres er avhengig av bindets forfatning.
Regesta Norvegica er ikke eldre enn at alle 7 bind er laget ved hjelp av datamaskinelle metoder. De må derfor ha eksistert i digitalt format. Men ikke alt av dette filmaterialet er bevart. Der filene er tapt, skannes boksidene inn; dette gjøres ved Kjeldeskriftavdelingen. Der materialet er bevart, har vi en del gamle, ukurante format som må konverteres; dette utføres av C.E. Ore.

5.2 Korrektur og koding

Neste trinn i prosessen utføres ved Kjeldeskriftavdelingen.
Det konverterte eller innskannede materialet blir korrekturlest. I tillegg koder vi teksten i en enkel, XML-basert koding. Vi bruker ikke spesialverktøy her, kun et standard tekstbehandlingsprogram.
Med kodingen merker vi ulike strukturelle enheter i teksten for å senere kunne avgrense søk til disse delene av teksten. Og vi merker enkelte spesifikke deler for å kunne søke direkte på dem. Vi skal komme tilbake til hvilke elementer det er snakk om.

5.3 Legge materialet over i base
Når vi så er ferdige med korrektur og koding, sender vi filene til Ore ved DOK-gruppen. På grunnlag av vår enkelt kodede fil, genereres en fullkodet fil som legges ut i basen.

6. Regesteksempel 2

6.1 Struktur og innhold i en regest
[ill. 8]
Vi går tilbake til regesteksemplet vårt for å se på hvilke deler av teksten vi koder i denne omgang.

Vi koder de tre feltene i overskriften: regestens nummer, dato og sted.
Så kommer selve regestteksten; denne tekststrukturen kodes. Det samme gjelder petitavsnittet, eller dokumentopplysningsavsnittet. Av rene strukturkoder er også fotnoter med fotnoteindikatorer.
Dermed vil vi kunne avgrense fritekstsøk til en av disse tekststrukturene. F.eks. vil vi kunne søke etter bestemte arkivsignaturer i dokumentopplysningsavsnittet.

I tillegg er der et par innholdselement som kodes.
For det første gjelder dette opplysningen om brevtype som kommer fremst i hver regest.
For det andre har vi interne regestnummerreferanser. Regester kan vise til andre regester; (et brev kan være svar på et annet). Slike interne henvisninger kodes for å muliggjøre koblinger i basen.
Tredje og siste punkt gjelder for brev trykt i Diplomatarium Norvegicum. Disse henvisningene kodes også slik at man får en kobling mot Diplomatariebasen.

6.2 Innholdselementer fanget opp av registrene
Man kan argumentere med at det er svært lite rent innholdsmessig som kodes i denne omgang. Det stemmer. Vi har bevisst plukket ut elementer som både er nyttige og som er raske å kode. Dessuten er hvert bind av det trykte Regesta utstyrt med et navne- og et sakregister, og i stedet for å kode hvert navn og hvert emne 'om igjen' så å si, har vi digitalisert de eksisterende registrene slik at de kan brukes som digitale oppslagsverk. Vi skal kort se hvor mye av innholdet i eksempelregesten vår som er fanget opp i registrene.
[ill. 9]
Person- og stedsnavn i navneregisteret er streket under med blått. Emner som gjenfinnes i sakregisteret er understreket med rødt.
[ill. 10]
Hvis vi nå slår alt dette sammen, ser vi at det i alt er ganske store deler av regesten som fanges opp i kombinasjonen av vår relativt enkle koding og de eksisterende registrene.

7. Demonstrasjon av Regestabasen
Dette er som sagt en prøvebase. Den består av de 3 første bindene i serien, (periode 822-1319) og inneholder ca. 3.400 regester.
På det stadiet vi er nå, er det ikke vesentlig hvordan ting ser ut, bare de virker, og skjermbildene kan til en viss grad bære preg av dette. Det kan forekomme enkelte uklare ledetekster i skjermbildene uten at disse vil bli nærmere kommentert her.

7.1 Søkebildet, 3-delt

[ill. 11]
Inngangsporten til Regestabasen er dette søkefeltbildet. Her kan man søke på ulike nivåer. For det første kan man slå opp i basen på samme måte som man kan slå opp i bøkene, direkte på et regestnummer eller et sidetall. Dette er nyttig dersom man har en henvisning direkte til regestnummer.
For det andre kan man søke direkte på de innholdselementene som er kodet; *sted, dato* og *brevtype*.
For det tredje kan man utføre fritekstsøk i de delene av teksten som vi kodet som strukturelle enheter. Man kan selvsagt også kjøre fritekstsøk i hele teksten.

7.2 'Vis regest'-bildet med link til DN

[ill. 12-14]
Vi begynner med å slå opp på den regesten vi brukte som eksempel tidligere; bind 3 nr. 467.

Her er vi i selve regestbildet. I tillegg til at vi her selvsagt får opp regesten, har vi et par manøvreringsmuligheter. Vi kan bla til forrige og neste regest i trykkrekkefølgen i boka. Videre ser vi her de interne regestnummerreferansene som ble kodet. Disse vil bringe oss direkte til den henviste regesten. Det skal vi ikke prøve nå, derimot skal vi prøve linken til Diplomatariebasen. Dette er jo en regest basert på et brev som er trykt i Diplomatarium Norvegicum. Og hvis vi klikker på DN-referansen, får vi opp Diplomatariebasen og det aktuelle brevet i et nytt vindu med Regestabasen liggende i bakgrunnen.

7.3 Søke på sted - treffliste

[ill. 15-16]
Vi går tilbake til søkebildet. La oss tenke oss at vi ikke kjenner regestnummeret, men tror brevet vi leter etter ble skrevet på Avaldsnes. Da søker vi på dette stedet, og får opp en treffliste hvor vi kan velge aktuell regest ut fra lista.

8. Registrene

[ill. 17-20]
En annen innfallsvinkel til søk i basen er å slå opp i registrene. Det ligger en link til registrene øverst i søkefeltbildet. Regesteksemplet vårt var et brev ført i pennen av 'Balte klerk', og nå skal vi undersøke om Balte har skrevet flere brev i basen.
I tillegg til Navneregisteret og Sakregisteret har vi laget et Felles navne- og sakregister for alle bind i basen, og vi skal gå inn her.
Foreløpig har vi ikke muligheten til å skrive inn det vi søker etter. Istedet får vi opp alfabetet og kan slå opp på ønsket bokstav; vi går da inn på 'B' for 'Balte'. Så får vi listet opp alfabetisk alle oppslagsord på 'B', og kan skrolle oss ned i lista til vi finner det vi leter etter.
Når vi nå velger 'Balte', får vi opp en treffliste av samme type som ved søk i søkefeltbildet. Lista viser at vi kun har 3 brev fra Baltes hånd i basen. Fra trefflista kan man som vanlig gå inn på ønsket regest.

9. Inkonsistensene

Så langt har vi sett på relativt enkle oppslag i basen, oppslag der vi finner det vi leter etter. Det gjør vi ikke alltid, noe som bl.a. skyldes at det i Regesta finnes en del inkonsistenser f.eks. i terminologibruk, mellom de enkelte bindene.

Regesta Norvegica er et verk som er blitt til over lang tid og under flere redaktører, så inkonsistenser er neppe til å unngå. Noen av forskjellene mellom bindene er bevisste redaksjonelle valg, andre er mer tilfeldige. Men uansett grunn får inkonsistensene konsekvenser for søkbarhet i en databaseversjon av verket. Vi skal se noen eksempler på det, nærmere bestemt ved valg av målform, ved typografisk oppsett og ved normalisering av patronymer.

9.1 To målformer; bokmål og nynorsk

Vi nordmenn er velsignet med 2 målformer, bokmål og nynorsk. Og i et nasjonalt verk av typen Regesta Norvegica er det selvsagt politisk korrekt å utgi noen bind på nynorsk og andre på bokmål. Dette får konsekvenser for oss når vi skal søke i basen. Prøvebasen inneholder ett nynorskbind (bind 2) og to bokmålsbind (bind 1 og 3).
[ill. 22]
Eksempelregesten vår var av typen 'Gavebrev'. La oss si at vi er interessert i å finne alle gavebrevene i basen. Da søker vi etter brevtype = 'Gavebrev', og får 76 treff. Men da har vi bare fått med forekomstene i bokmålsbindene. I nynorskbindene heter samme brevtype nemlig ’Gåvebrev’. Og søker vi på denne brevtypen, finner vi 34 regester til. Dette er ikke en akseptabel situasjon. Vi har såvidt diskutert muligheten for å lage en brevtype-ordliste og koble sammen alle termer for samme brevtype slik at når man søker etter 'Gavebrev' vil programmet ta med 'Gåvebrev' også. Men i denne fasen av prosjektet er vi i forhold til inkonsistensene foreløpig i en kartleggingsfase; vi peker på problemene, løsningene får komme senere når vi har alle dataene inne i basen og får oversikt over omfanget av det enkelte fenomen.

Foreløpig er det måter å unngå problemet på. Vi kan kjøre fritekstsøk og bruke logiske operatorer eller jokertegn i søkeuttrykkene.
[ill. 23]
Her har jeg søkt etter forekomst av teksten 'Gavebrev' eller 'Gåvebrev', (søkeuttrykk: *Gavebrev|Gåvebrev*), og resultatet av dette søket var 120 regester. Her er noe som ikke stemmer. Hvis vi går tilbake til de to separate søkene var det 76 + 34 hvilket til sammen blir 110 og ikke 120 regester. Det er en enkel forklaring på dette. I det første tilfellet søkte jeg etter forekomster av 'Gavebrev' og deretter 'Gåvebrev' i brevtypefeltet. I det andre tilfellet søkte jeg etter de samme termene i hele regestteksten. Og det er klart at ord som gavebrev/gåvebrev kan forekomme i regestene også utenom selve brevtypeklassifiseringen, og slike forekomster vil komme med i fritekstsøket. Dette er en generell tendens når man kjører fritekstsøk med logiske operatorer i motsetning til det vi kan kalle ferdigdefinerte søk; man får ofte med mer enn man trodde man søkte etter. Og det er det viktig å være klar over.

9.2 Typografi i registrene

Inkonsistenser i typografisk oppsett er særlig tydelig i registrene. Følgende eksempel kombinerer variasjon i typografi og valg av målform.
[ill. 24]
Her har jeg søkt i fellesregisteret etter Kristkirken i Bergen. Og det viste seg at hvert av basens 3 bind hadde sin egen utforming av dette oppslagsordet. Ergo ble det tre oppslag i registeret istedet for ett.

Bind 1 som er et bokmålsbind har 'Kristkirken' og stedsangivelse i skarpe klammer.
Bind 3 er også bokmålsbind, men her er sted angitt med preposisjonsuttrykk.
Bind 2 er nynorskbind og har formen 'Kristkyrkja' og sted som preposisjonsuttrykk.

Jeg kan tenke meg minst én variant til her etterhvert som flere bind blir lagt inn i basen.

Her skal det kun være ett oppslagsord; det er samme kirken i alle tilfellene. Men å slå sammen ulike former for samme oppslagsord blir en jobb som først kan gjøres når alle binds registre er på plass i basen.

9.3 Patronymer på vokal

[ill. 25-27]
Regesta normaliserer navn til moderne norsk form. Vi skal ta for oss patronymer, nærmere bestemt patronymer der farsnavnet ender på vokal.

Patronymer generelt er som regel gjengitt med dobbel s i Regesta, dvs. eieformens/genitivens -s + ’son’ som i eksemplet her: Olavs + son = Olavsson

Patronymer der farsnavnet ender på vokal har historisk sett ikke hatt dobbel s; - her var det opprinnelig ingen s-ending i eieformen, - og noen redaktører av Regestabind har derfor valgt å gjengi patronymer på vokal med enkel s. Dette gjelder bl.a. i bind 1. Andre redaktører lar alle patronymer, uansett historisk form, få dobbel s. Dette gjelder for de fleste bind av Regesta, bl.a. for bind 2 og 3 i prøvebasen.

Dette får konsekvenser for søking i basen. Igjen kan vi bruke jokertegn i fritekstsøk for å unngå problemet: Vi søker etter *Arnes%on* i regestteksten. %-tegnet betyr '0 eller flere tegn'. Dette fritekstsøket vil finne alle forekomster av 'Arneson' med 1 eller 2 s’er, men det kan finne mye annet også, f.eks. en tekstsekvens som ’Arnes bataljon’. Så ved slike søk er det viktig å sjekke resultatet.

Resultatet er trefflisten med 42 treff. Når trefflisten er lang er den i tilfeller som dette ikke så informativ. Vi har derfor muligheten til å vise selve regestteksten allerede i trefflista, slik at vi raskt kan sjekke hvorvidt vi faktisk har fått treff på en forekomst av patronymet 'Arnes(s)on' og ikke noe helt annet.

10. Avslutning

Det finnes flere typer inkonsistenser, (navn er et stort emne, og datofeltene er et kapittel for seg), men jeg tror de eksemplene jeg har vist er tilstrekkelig til å illustrere hvilke utfordringer vi står overfor.

I prosjektets fase 1 må vi kartlegge alle slike tilfeller av variasjon, slik at vi i den ferdige basen kan få søkefunksjoner som gir brukeren fornuftige resultater samtidig som søkefunksjonene skal være enkle å bruke.
I hvilken grad vi klarer å realisere dette og hvilke løsninger vi velger, det vil tiden vise.

Takk for oppmerksomheten.