Datamodellering af geografiske data

Anders Friis-Christensen, Kort & Matrikelstyrelsen

Introduktion
Forskning inden for datamodellering af geografiske data er nødvendig af flere årsager. Dels er brugen af geografisk information og geografiske informations systemer (GIS) i stadig stigning. Dels vokser behovet for at udveksle heterogene data mellem forskellige institutioner og myndigheder. Et eksempel på heterogene data kunne i denne forbindelse være data, der er indsamlet og organiseret forskelligt, men beskriver det samme 'virkelighedens begreb', som f. eks. bygningerne i tekniske kort og i TOP10DK. Typisk er disse data indsamlet og organiseret forskelligt på baggrund af forskellige, ikke standardiserede datamodeller, som igen er applikationsafhængige. Dette gør integrationen og udvekslingen af data vanskelig, og en standardiseret måde at strukturere og formalisere data på vil gøre integrationsarbejdet betydeligt nemmere.

Geografiske data har mange fælles egenskaber (f.eks. tid og sted), der ikke understøttes i modelleringssprog, som f.eks. Entitet/Relations-diagrammer (E/R-diagrammer) og *Unified Modeling Language* (UML). Da disse egenskaber ikke er understøttet, opstår risikoen for at samme begreb bliver beskrevet, tolket og implementeret på flere forskellige måder.

Denne artikel tager udgangspunkt i mit ph.d. studium, og beskriver hvilke fælles egenskaber ved geografiske data man med fordel kan modellere. Eksempler på hvordan eksisterende metoder kan benyttes i en modellering af geografiske data præsenteres, og til sidst identificeres områder hvor der er behov for en videreudvikling.

I artiklen, såvel som i mit ph.d. studium, har jeg en objektorienteret tilgang til datamodelleringen, og UML benyttes som modelleringssprog. UML er valgt, fordi det er et meget benyttet modelleringssprog inden for objektorienteret systemudvikling både i industrien og i forskningsmiljøerne. UML benyttes endvidere i standardiseringsarbejdet omkring geografisk information, ISO-TC211 (http://www.statkart.no/isotc211).

Definition af geografisk datamodellering
Datamodellering bruges i mange forskellige sammenhænge, og jeg skal her komme med en definition i forhold til mit arbejde. En datamodel er en model for, hvordan virkeligheden repræsenteres og lagres i en database. Datamodelleringen er processen, der skaber en datamodel. Geografisk datamodellering er datamodellering, der beskæftiger sig med geografiske data, som er kendetegnet ved tre basale komponenter: sted, tid og tema.

Der findes flere niveauer af datamodellering. I databaselitteraturen opereres generelt med 3 niveauer: en konceptuel, en logisk og en fysisk model.

Den konceptuelle model er idealiseringen af den verden, vi ønsker at beskrive i en bestemt kontekst og til et bestemt brug. De konceptuelle modeller beskriver hvilke objektklasser der findes og eventuelle relationer mellem dem. Eksempler på modelleringssprog på dette niveau er E/R-diagrammer og UML.

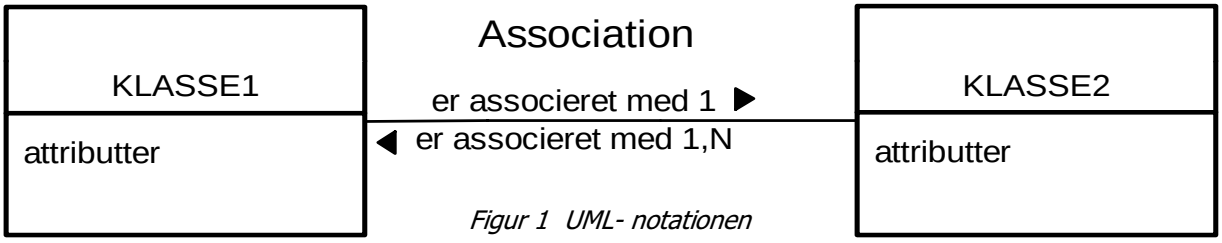
Den logiske model forbereder den konceptuelle model til implementering, eksempelvis til den relationelle model, som kan implementeres i alle relationelle databasesystemer (f.eks. Oracle eller Informix).

Den fysiske model er den konkrete model i computeren, og den beskriver f.eks. indekseringer, samt hvordan data ligger lagret.

Da konceptuelle modeller er kontekstafhængige, kan samme virkelighedens begreb repræsenteres flere gange alt afhængig af formålet. For at håndtere dette aspekt kan man indføre en domæne ontologi (Gaurino, 1997), som beskæftiger sig med overordnede begreber, f.eks. en bygning, og som kan benyttes i en beskrivelse af, hvordan de forskellige repræsentationer af det samme virkelighedens begreb er relateret til hinanden. Mit ph.d. studie beskæftiger sig primært med geografisk data-

modellering på det konceptuelle niveau, men det er også planen at inddrage arbejde med ontologibegrebet. I denne artikel vil det dog ikke blive beskrevet nærmere.

Som nævnt ovenfor benytter jeg UML. Notationen for klassediagrammer, som er anvendt i denne artikel, fremgår af Figur 1. Diagrammet viser f.eks., at et objekt af Klasse2 er associeret med et eller flere (1,N) objekter af Klasse1. Jeg vil ikke komme ind på en nærmere beskrivelse af UML, men blot henvise til f. eks. Booch et al. (1999) eller Fowler (1997).



Figur 1 UML- notationen

Behov for udvidelser af eksisterende metoder og teknikker

Som antydet i introduktionen er der et behov for at udvide UML's udtrykskraft i forbindelse med geografiske data. Dette kan gøres ved en konkret udvidelse af sproget, eller man kan skabe nogle mønstre eller skabeloner for, hvordan bestemte egenskaber skal modelleres. UML er meget generel, og behovet er hovedsageligt baseret på at eliminere den tvetydighed, der opstår i forbindelse med modellering på en ikke-standardiseret måde. Men et andet problem med UML er, at de konceptuelle modeller bliver meget komplekse, når de generelle egenskaber ved geografiske data skal modelleres (som f. eks. sted og tid), og de mister deres egentlige formål: at skabe et formelt ikke-teknisk overblik over de begreber man ønsker at lagre i en database.

Egenskaber ved geografiske data

Geografiske data har mange egenskaber, der er relevante at modellere. Følgende egenskaber er fundet ved bl.a. at gennemføre en behovsanalyse i Kort & Matrikelstyrelsen:

- spatiale og temporale egenskaber. Geografiske data har en udbredelse i tid og rum,

- associationer mellem geografiske objekter. Der findes flere typer af associationer mellem geografiske objekter, f.eks. de geometriske eller de topologiske. Den geometriske association mellem to objekter kan være afstand, og den topologiske association kan være 'overlapper', 'rører ved' osv.,

- multiple repræsentationer af geografiske objekter. En geografisk entitet (et virkelighedens objekt) kan være repræsenteret afhængig af formål og brug. Et eksempel kan være veje. I den matrikulære verden har veje en helt anden rolle end i den topografiske verden,

- regler for geografiske objekter. Regler er med til at opbygge den idealiserede verden vi ønsker at beskrive. F.eks. en regel der siger, at bygninger mindre end 20 m^2 ikke skal medtages i vores idealiserede verden.

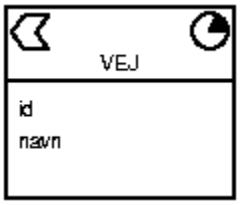
Ovenstående er ikke en komplet beskrivelse af egenskaber. Se Friis-Christensen et al. (2001) for en mere uddybende beskrivelse.

Teknikker og metoder til modellering

Følgende afsnit beskæftiger sig med konkrete modelleringsmetoder og beskriver mere detaljeret hvordan de enkelte egenskaber kan modelleres, samt hvilke krav de enkelte egenskaber stiller til udvikling af modelleringsmetoder.

Spatiale og temporale egenskaber og relationer

De spatiale egenskaber dækker over geometri, som enten kan være 0-,1-,2- eller 3-dimensional (punkt, line, polygon eller volumen). De temporale egenskaber dækker over valid tid, dvs. hvornår et objekt er gyldigt, samt transaktionstid, dvs. hvornår der er sket ændringer med objektet i databasen. Herved kan historikken af objektet lagres.



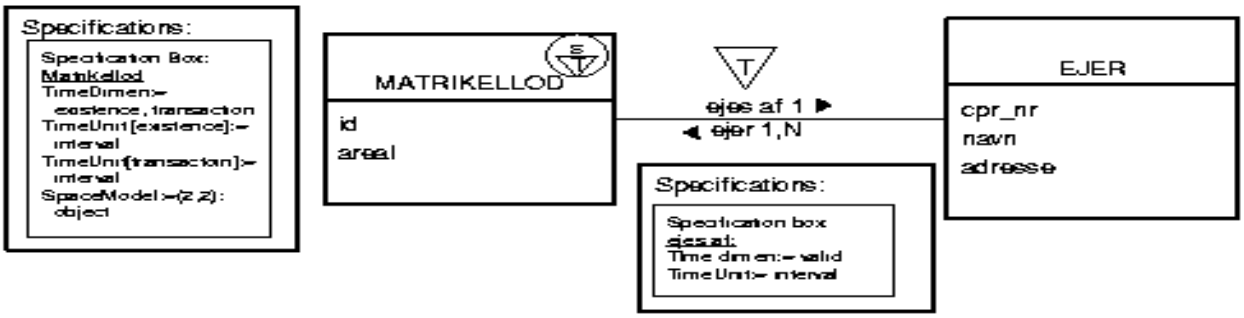
Figur 2 Eksempel på en vej modelleret i Perceptory

Udvidelser af UML med denne type egenskaber er bl.a. gjort i et *Case Tool* kaldet *Perceptory* (Proulx og Bédard, 2001). Et eksempel på en udvidelse er vist på ovenstående Figur 2.

Her er UML udvidet med nogle såkaldte stereotyper, som blot er nye specialiserede komponenter, der udvider en generel klasse. Stereotyperne beskriver i dette tilfælde de spatiale og temporale egenskaber: at en vej er en polygon og eksisterer (er valid) over en given periode.

Perceptory kan ikke udtrykke alle former for spatiale og temporale egenskaber. F.eks. kan man ikke udtrykke, at et objekt har en 3-dimensional geometri og en transaktionstid. Ligeledes kan Perceptory ikke udtrykke specielle geografiske relationer mellem objekter, som f.eks. en topologisk afhængighed mellem objekter.

Et eksempel på en notation der er mere udtryksfuld end Perceptory, er *Extended Spatiotemporal UML* - STUML (Price et al., 2000). STUML inkluderer både 3-dimensionale objekter, transaktionstid og visse former for geografiske relationer. Et eksempel kan ses på Figur 3, som viser at en *Matrikellod* ejes af en *Ejer*. En *Matrikellod* er et 2-dimensionalt geometrisk objekt, hvor vi registrerer hvornår objektet eksisterer og hvornår det ændres. Associationen *Ejes_af*, er en temporal association, da vi skal kunne registrere de forskellige ejere.



Figur 3 En matrikellod og en ejer modelleret i STUML. Specifikationsboksene kan skjules for at give et bedre overblik og gøre modellen mere simpel.

STUML er mere kompleks end Perceptory, men kan også udtrykke flere egenskaber ved geografiske data. Fordelen ved begge notationer er, at de kan oversættes direkte til et databaseskema. Perceptory understøtter endog direkte Oracles geometriske datatyper (baseret på *Open GIS Consortium*).

Udover de nævnte findes en række andre notationer med hver deres fordele og ulemper, f.eks. MADS (Parent et al., 1999) eller GeoFrame (Filho and Iochpe, 1999). En direkte anvendelse af et værktøj som understøtter geografiske data, kræver en tilpasning til et konkret brug, f.eks. som en kombination af de eksisterende notationer. Selve grundlaget for denne udvikling er dog til stede.

Multiple repræsentationer af geografiske objekter

I modsætning til de spatiale og temporale egenskaber eksisterer der ikke nær så mange forskningsresultater inden for konceptuel modellering af eksempelvis multiple repræsentationer og regler for geografiske objekter. Multiple repræsentationer af geografiske objekter inkluderer en række forskellige områder som f.eks. generalisering, skala og kontekstafhængig repræsentation. Multipel repræsentation opstår i de tilfælde, hvor der findes flere kontekstafhængige repræsentationer af det samme virkelighedens begreb. Det optræder hyppigt i GIS-verdenen, f.eks. kan en vej være repræsenteret i flere forskellige skalaer, eller en bygning kan være repræsenteret i et register eller som et grafisk element i en kortdatabase. Multiple repræsentationer afhænger derfor af formål med repræsentationen og metode til dataindsamling osv. For en uddybende beskrivelse af multiple repræsentationer i geodatabaser kan bl.a. henvises til Kilpeläinen (1997). Problemet med multiple repræsentationer er, at det skaber inkonsistens imellem vores repræsentationer af det samme begreb, da der typisk ikke er relationer imellem repræsentationerne. Der findes en del forskningsaktiviteter inden for automatisk generalisering af geoda-

tabaser, hvorved et objekt i én skala er afledt af et objekt i en anden skala. Her kan den specifikke relation mellem disse to objekter lagres. Der er dog brug for flere metoder til at håndtere relationer mellem multiple repræsentationer, og ydermere et behov for at kunne understøtte dem med en notation i en konceptuel model.

Regler for geografiske objekter
Regler for geografiske objekter er vigtige, da det er dem der specificerer, hvordan vi ønsker at repræsentere vores idealiserede verden. Der findes flere forskellige inddelinger, som er tæt knyttet til relationerne mellem objekter: de topologiske (f.eks. 'overlapper' og 'indeholder'), de geometriske (f.eks. afstand) og de tidsmæssige regler (f.eks. synkroniseret). Desuden er der regler, der knytter sig til selve objekterne, f.eks. størrelse, tilladte værdier på attributniveau osv. Er en regel ikke overholdt, fejler den, og et objekt indsættes ikke i en database. I UML findes et regelsprog *(Object Constraint Language),* som kan benyttes til at specificere regler. Sproget er dog ikke videre brugervenligt, og desuden indeholder det ikke definitioner af specifikke operatorer, der relaterer sig til geografiske data. Der findes f.eks. ikke en operator, der kan specificere, at et objekts geometri ikke må krydse et andet objekts geometri. Derfor er der behov for en udvidelse af dette regelsprog, så det understøtter geografiske data.

Konklusion og videre arbejde
I artiklen er givet eksempler på forskellige aspekter af geografisk datamodellering. Samtidig fremgår det i hvilke områder, der allerede forefindes en del forskningsresultater, og i hvilke der stadig mangler forskning. Der findes flere specifikke notationer for modellering af spatiale og temporale egenskaber af objekter og relationer. Hvilken notation der skal bruges, eller en eventuel videreudvikling, afhænger af den givne anvendelse. Hvis en notation tages i brug, vil det skabe grundlag for konsistens mellem forskellige modeller, og samtidig kan fokus i selve modelleringsprocessen lægges på modellering af begreberne i stedet for på geometri- og tidsmæssige aspekter. Resultatet vil være en konceptuel model, der forholdsvis simpelt kan implementeres i en database.

Mit ph.d.-studie beskæftiger sig primært med de områder, der ikke har været i fokus, dvs. konceptuel modellering af multiple repræsentationer og regler for geografiske data. Ydermere er et aspekt som kvalitet af geografiske data også yderst relevant. Det er endnu et forholdsvist uudforsket området, især inden for modellering af datakvalitet.

Referencer

Booch, G., Rumbaugh, J. and Jacobsen, I. (1999). *The Unified Modeling Language user guide*, Addison-Wesley, USA.

Filho, J.L. and Iochpe, C. (1999*). Specifying Analysis Patterns for Geographic Databases on the basis of a Conceptual Framework*, Proceedings of ACM GIS 1999, Kansas City, USA.

Fowler, M. (1997). *UML Distilled: Applying the Standard Object Modeling Language*, Addison-Wesley, USA.

Friis-Christensen, A., Tryfona, N. and Jensen, C.S. (2001). *Requirements and Research Issues in Geographic Data Modeling*, Proceedings of ACM GIS 2001, Atlanta, USA.

Gaurino, N. (1997). *Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration. In M. T. Pazienza (ed.) Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Springer Verlag: 139-170.

Kilpeläinen, T. (1997). *Multiple Representation and Generalization of geo-databases for topographic maps*, Publications of the Finish Geodetic Institute, no. 124.

Parent, C., Spaccapietra, S. og Zimányi, E. (1999). *Spatio-temporal conceptual models: Data structures + space + time*, Proceedings of ACM GIS 1999, Kansas City, USA.

Proulx, M.-J. og Bédard, Y. (2001). *Perceptory, Centre de recherche en geomatique*, Université Laval, Quebec, Canada, http://sirs.scg.ulaval.ca/perceptory.

Om forfatteren

Anders Friis-Christensen, Ph.D-studerende, Datalogisk Institut, Aalborg Universitet
Arbejde: Kort & Matrikelstyrelsen, Rentemestervej 8, 2400 København NV
e-mail: afc@kms.dk