MODELLING A DYNAMIC GEODATA BASE: PROBLEMS OF DATA ACCURACY AND STRUCTURE CONVERSIONS IN DATA COLLECTION AND PROCESSING

Kirsi Makkonen (formerly Eloranta)
Helsinki University of Technology, Institute of Geodesy
Espoo, Finland
Comission IV

ABSTRACT
This paper deals with dynamic aspects of geodatabase modelling. Especially the problems of data accuracy definition and data structure conversions are treated. A taxonomy of accuracy concepts as well as some new notations for logical data structure descriptions are proposed.

1. INTRODUCTION

Land information systems (LIS) and their data bases differ from the "ordinary", administrative information systems in the spatial dimension of entities. Information in a LIS are land related - and the land is unique - thus several references are associated with the same piece of land. In addition to this, land related information processing activities are often subprocesses of larger systems and thus "chained" procedures. Most land information systems, however, do not reflect the integration of their object systems. LIS's are rather implemented as separate systems, to serve e.g. cadastre, land use planning, mapping, natural resource management, utility survey, city engineering, building etc. Data bases are isolated models of one application and thus often typically static descriptions of entities and relations.

However, data descriptions of a LIS, especially of an integrated LIS, are not time-invariant. Activity integration includes, in addition to spatial relationships, also temporal dependencies. Chrisman (1983) describes the dynamics and its complexity in a LIS by saying: "Space, time and attributes all interact." We call this interaction as a process of entity evolution. The word "evolution" is used to emphasize the process of changing from one entity to another inside the data base.

By a change in dynamic modelling is ment in general (ISO, 1980):
"- entities appearing or disappearing from the object system,
- an entity changing its properties or associations with other entities,
- the classification of subjects or some rules in the object system changing as a result of policy changes,
- the scope of interest changing, so that the object system itself expands or shrinks."

In geometric representation of entities the changes mainly appear in data accuracy and data structures. Geodatabase dynamics is approached in this paper by outlining some practical entity evolution processes and by analysing their data accuracy and structure changes. Some new accuracy concepts are introduced and the notation of conceptual description of data structures is broadened.

2. DATA ACCURACY

2.1 Problems with data accuracy definitions

It is an amazing effect which the digital/numerical information have on the users. Almost anything retrieved from computer memory and expressed in numericals instead of verbal or graphical descriptions seems "accurate" and "liable". Thus the potential unintended or intended misuse is obvious. Specifications of data accuracy - both of quality and quantity - should be associated with all entities of data base. It is a part of the so-called "quality-information" of data (Chrisman, 1983).

In data collection data accuracy depends both on the registration method and performance as well as on the source material. We can distinguish between simple registration methods (like digitizing of contours on a line map) and registration with interpretation (stereoplotting is an example of this type). In simple registration the operator only has to try to follow the line as accurately as possible. Errors are caused by hardware or software limitations and the careless work of the operator. In registration with interpretation the information have to be e.g. classified, so skill and experience are needed.

Material can also be classified according to the quality into two groups: exact and fuzzy material. Exact material, like drawn line maps, can be registered without interpretation. Stereomodels are examples of not always "clear" material. Boundaries to be registered do not exist as lines, but there rather is "a fuzzy zone of interpretation and transition" (Chrisman, 1982).

The errors in the cases above are caused by different factors and thus the data accuracies can not be directly compared with each others.

When dealing with data accuracy during data processing we have to distinguish between two types of processes: "closed" processes with predefined procedures and accuracy changes and processes including interaction with casual information inputs, data comparisons and development. Coordinate transformations and automatic generalization are examples of closed processes. Accuracies of data sets can be well defined because they are well known in the stage of programming, and at least at the parameter input phase. Drafting is an activity in which data accuracy can not only decrease or remain as it was but even be improved because of the information input of the user. Accuracy seems thus to have here also the temporal dimension.

In geodata bases map information must be stored for accuracy. On the map the same information should be presented for visualization. This statement is fundamental for accuracy analysis in map production and in land information systems - yet in a critical way: data accuracy itself has no absolute value, the use of information states the relevant requirements.

2.2 Accuracy types

In order to discuss data accuracy in a LIS we have to define the used concepts. As we saw above all errors can not be compared because of their different nature. In the following we introduce one taxonomy of useful accuracy concepts (Figure 1).

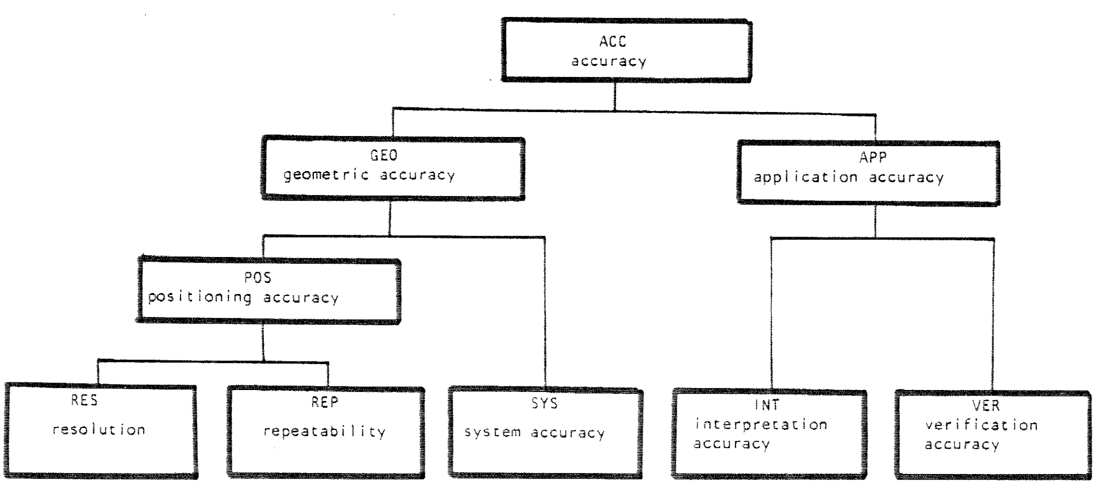

Fig. 1. Accuracy types.

The most common accuracy concepts associated with geodata collection, processing and outputting are resolution and repeatability. Generally accepted definitions of them can be found in literature. According to (Deumlich, 1982) and (SFS, 1982) we can formulate the following definitions:

- Resolution (RES) of a system indicates its capability of imaging two neighbouring object points separately.
- Repeatability (REP) is the closeness of the agreement between the results of successive measurements of the same quantity carried out by the same method, by the same observer, with the same measuring instruments, in the same environment, at quite short intervals of time.

Resolution and repeatability make the positioning accuracy (POS). Positioning accuracy then includes the technical (hardware and software) aspects of the accuracy.

Positioning accuracy does not, however, express the realistic total error in the entire data registration process. The term, execution accuracy (Wolfendale, 1967) is sometimes used to describe the positioning accuracy accomplished with a human performance factor. In this paper we introduce another term, namely geometric accuracy (GEO) to emphasize the difference between errors associated with geometric (boundary) information and errors associated with the semantics of the map to be registered.

System errors - system accuracy (SYS) - include errors caused by human performance, actual circumstances, ergonomy etc., as well as deformation of the material. By human performance we mean here the ability (carefulness) of the operator e.g. to follow a given line.

When the human performance includes also decisions e.g. about the type/class of registered objects, the operation is said to contain interpretation. In this kind of registration procedures the interpretation accuracy (INT) is an essential variable. Interpretation error is defined to be the error which occurs when the operator has to identify the object type by using his/her experience. In interpretation there are two factors: the operator and the source material. Interpretation errors can be omitted if the registered results can be compared with the real situation.

All material to be registered, processed or output can not be compared with the original source because sometimes information with no real counterparts have to be processed. Plan maps are examples of this kind of information. Associated with this material we can speak about certain (planning) tolerances within which the boundaries can vary without distorting the information contents of the map. Each map naturally has this kind of tolerance but in the case e.g. plan maps the tolerance is relatively high. This accuracy type we call here as verification accuracy (VER).

Interpretation accuracy and verification accuracy together make application accuracy (APP).

Among two (or more) data sets one of them can be considered as the reference data set and thus also the accuracy of it is called as reference (REF) accuracy. As the opposite to the reference data accuracy the other data sets are said to have relative (REL) accuracy. Coordinate values of a reference data set are used in transformations as fixed points and the other data sets are adjusted according to it. A data set with reference accuracy, in a way or another, bounds the other data sets with "local" relative accuracy to a larger, more "global" context. An example about these accuracy types can be taken from stereoplotting; data of one model have relative accuracy and the fixed points which are used in the block adjustment (geodetic points) are considered as reference information with reference accuracy. On the other hand the same geodetic points have relative accuracy when the question is about national measurements. Thus reference accuracy is not a fixed characteristic but depends always on the viewpoint and the application.

Two data sets can also be combined without naming either of them as reference information. Neither of them is then more "global" than the other. E.g. when two draft plans are put together coordinate transformations are based on the grid crosses of the base map, but the relative nature of the accuracy still remains.

3. DATA STRUCTURE CONVERSIONS

3.1 Data structure conversions as a part of the application

As stated the dynamics of a geodata base has two factors:

- changing data accuracy and
- data structure conversions.

These change processes are placed in a generalized diagram describing computer aided map producing process (see Figure 2.)

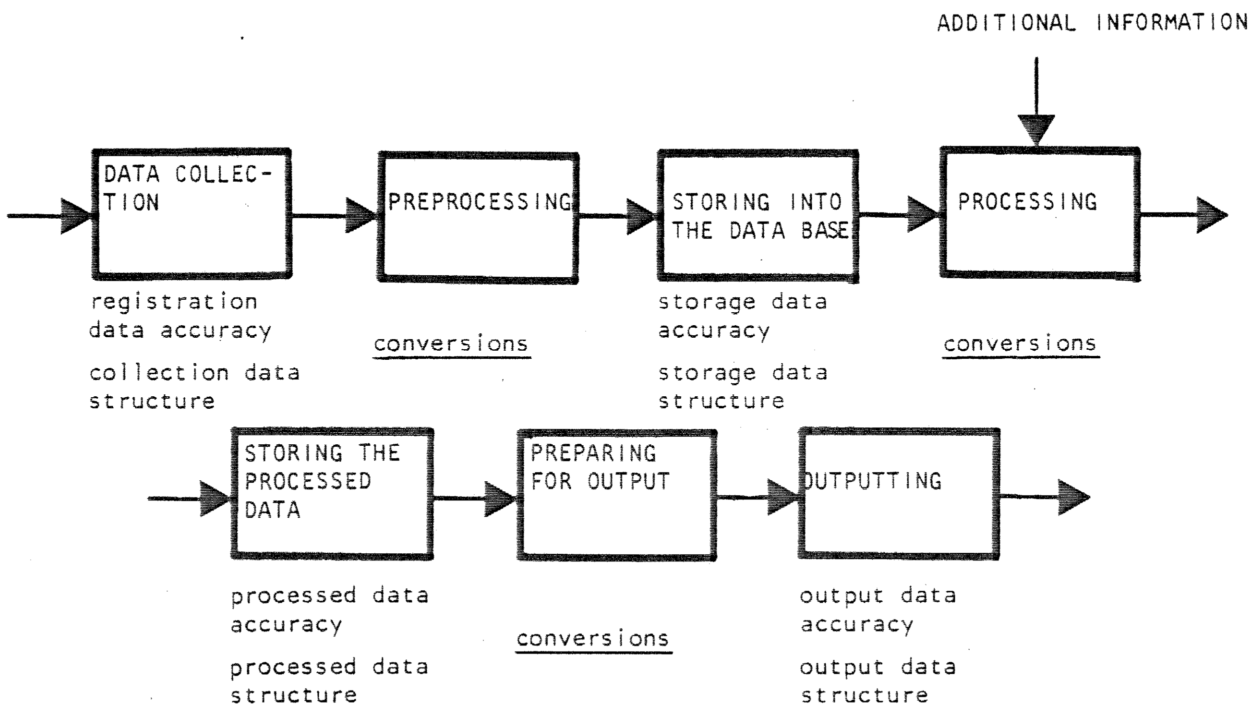

Fig. 2. A generalized diagram describing map producing process supported by computer. All data processing stages have specified data structures and data accuracy. Between the stages certain changes will happen.

Each data collecting, processing and outputting technique prefers certain data structure as well as it is associated with a certain type and level of data accuracy. Scanners produce matrices; digitizing by manual digitizing tables creates points, lines or polygons depending on the selected logic; stereoplotters produce typically segments or points which in a traditional procedure are not even coded but only registered in an arbitrary manner; field surveys produce information in numerical documents. All of these information must be combined and stored in a data base, efficiently and easily. In the previous chapters we already discussed the problems of data accuracy definition. Finding the right data structure requires at least as thorough a semantic analysis of the data contents and the use of information.

Data structure conversions can be distinguished in a LIS on three levels: conversions of entire data bases, files or records. In all cases the conversions should be seen as functional parts of the information system and thus they should also be seen in the data base descriptions, the conceptual schemas. In the following we deal with the descriptions of the conversions, first introducing the general approach to geometric data representation based on the EAR -approach (ISO, 1980).

3.2 On describing data structure conversions

Geometric representations of entities

Entities of a LIS (buildings, streets, parcels, blocks, terrain features etc.) are either geometric or non-geometric depending on whether the geometric representation of their outlines or image can be retrieved from the system. In a linked organization [1] the outlines of an entity can be given:

- explicitly by coordinate values,
- implicitly by pointers to other entities which are defined explicitly,
- as a deferred representation by expressing the location of a non-geometric entity by a reference to a geometric entity.

Geometric representations are based on the following geometries:

- pointgeometry: a triple (x,y,z),
- linegeometry: an ordered set of (x,y,z) triples,
- areageometry: a closed linegeometry with certain additional characteristics.

For more detailed definitions of the concepts see (Eloranta, K., 1982).

Four relations have been defined:

EXPLICITGEOMETRY (between entities and geometries),
IMPLICITGEOMETRY (between geometric entities),
DEFERREDGEOMETRY (between geometric and non-geometric entities) and
BELONGSTO (between entities and entity classes or groups).

BELONGSTO is used for describing partitions (entity classes, groups, networks etc.).

In addition to the previously mentioned geometric relations - EXPLICITGEOMETRY, IMPLICITGEOMETRY and DEFERREDGEOMETRY - lots of various geometric relations can be used for entity descriptions. In Chapter 4 we meet some examples of them ("on-the-right", "on-the-left") which are succesfully used e.g. in information registration. However, in order to utilize the polygon outline e.g. in area computations, plotting etc. the geometric representation should be reduced into either implicit or explicit form. The examples in Chapter 4 also show the need for these structure conversions.

1) Two main approaches can be distinguished associated with the internal data organization in a LIS: cellular organization and linked organization (Nagy et al., 1979).

Describing the dynamics

In dynamic modelling of a data base static models are linked into a chain of sequential representations. The dynamics can be described e.g. by a state diagram (ISO, 1980). In this paper we give a proposal about a more detailed description of dynamics based on conceptual schemas.

In order to describe conversions in dynamic schemas we have to define two new relations between entities:

CONSISTENTWITH and
TRANSFORMEDFROM.

For CONSISTENTWITH we have to define two auxiliary concepts: geometric hierarchy and geometric dependency.

Geometric hierarchy means hierarchic relation between one or more geometric entities in the sense of IMPLICITGEOMETRY.
Geometric dependency exists between two or more geometric entities whose geometric representations depend on each others in the following way: creation/transformation/deletion of the other causes creation/transformation/deletion of the another; however, these two entities must exist at the same time in the same information base.

CONSISTENWITH exists between two or more geometric entities whose geometric representations have geometric dependency between each others but which are not geometrically hierarchic. Entities in CONSISTENWITH relationship exist at the same time in the same information base.

TRANSFORMEDFROM exists between two or more entities which belong to separate conceptual schemas describing sequential states of entity evolution. TRANSFORMEDFROM represents the transformation which occurs between data structures including the needed interpretation and semantic change.

The following notations are used in conceptual schemas (see Figure 3).

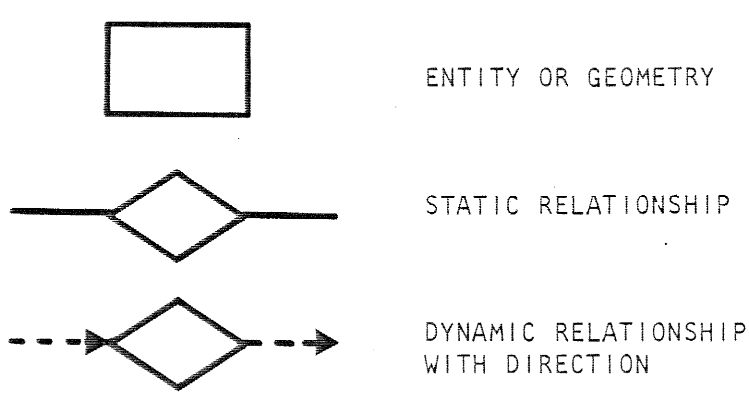

Fig. 3. Notations in conceptual schemas.

4. TWO EXAMPLES OF ENTITY EVOLUTION IN A LIS

4.1 Forestry taxing

In Finland forest owners are taxated according to the area and the quality of the forest in question. In order to calculate the amount of taxes a classification of forest types has been generated. The task in forestry taxation is to identify the different forest type units inside each real estate and to compute their areas as well as the final amount of taxes by using a certain weighting procedure and formula.

The original material consists of

- aerial photographs for stereodigitizing the boundaries of forest units,
- graphical cadastral maps for tabledigitizing of real estate boundaries.

The registration logics can be derived from the schemas (in Figure 4):

- In A the forest type boundaries are registered as segments in arbitrary order, plus one reference point inside each forest type area.
- Later on in processing the polygons are created by concatenating the successive lines which make the smallest polygon around the reference point (schema C).
- In B the real estate boundaries are registered by using the topological right-left-coding.
- The direction of registration is also stored and closed polygons can be created by sorting the segments (schema D).
- When these two data sets are registered and polygons created the polygon overlays are computed in order to specify the forest type subunits in each real estate. The final data structure is shown in schema E.

In addition to data structure conversions also data accuracies change during the processing - although they change according to predefined procedures because the question is of a closed process. Digitized coordinates from the cadastral map are chosen to be the reference information and thus they also have reference accuracy. Tabledigitizing of cadastral maps is mostly simple registration of exact material and the geometric accuracy dominates. Stereodigitized material can be considered fuzzy and a certain amount of interpretation is needed - the accuracy type is relative and interpretation. The accuracy of the combined material is interpretation but reference. In the figure 4 the dominating accuracy types are shown.

4.2 Land use planning

The development process from a draft plan into the final city plan, site layout plan and real estate boundaries is a good example about an entity evolution process during which entities are changed into another entities. Data structure and accuracy conversions take place between (Figure 5.)

- digitized draft plan units (schema A),
- coordinated parcels of the site layout plan (schema B) and
- formated real estates (schema C).

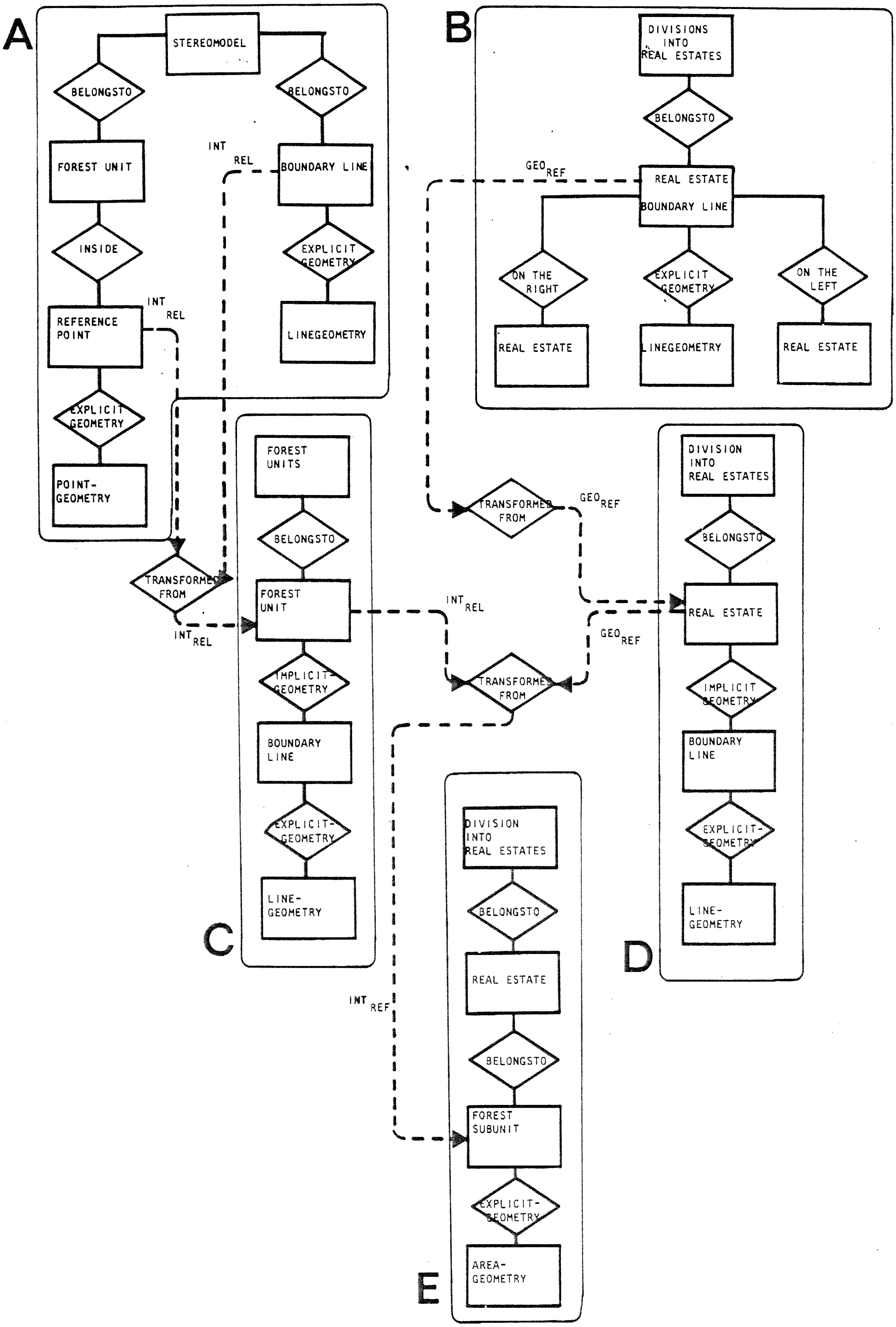

Fig. 4. Dynamic conceptual schema about forest taxation.

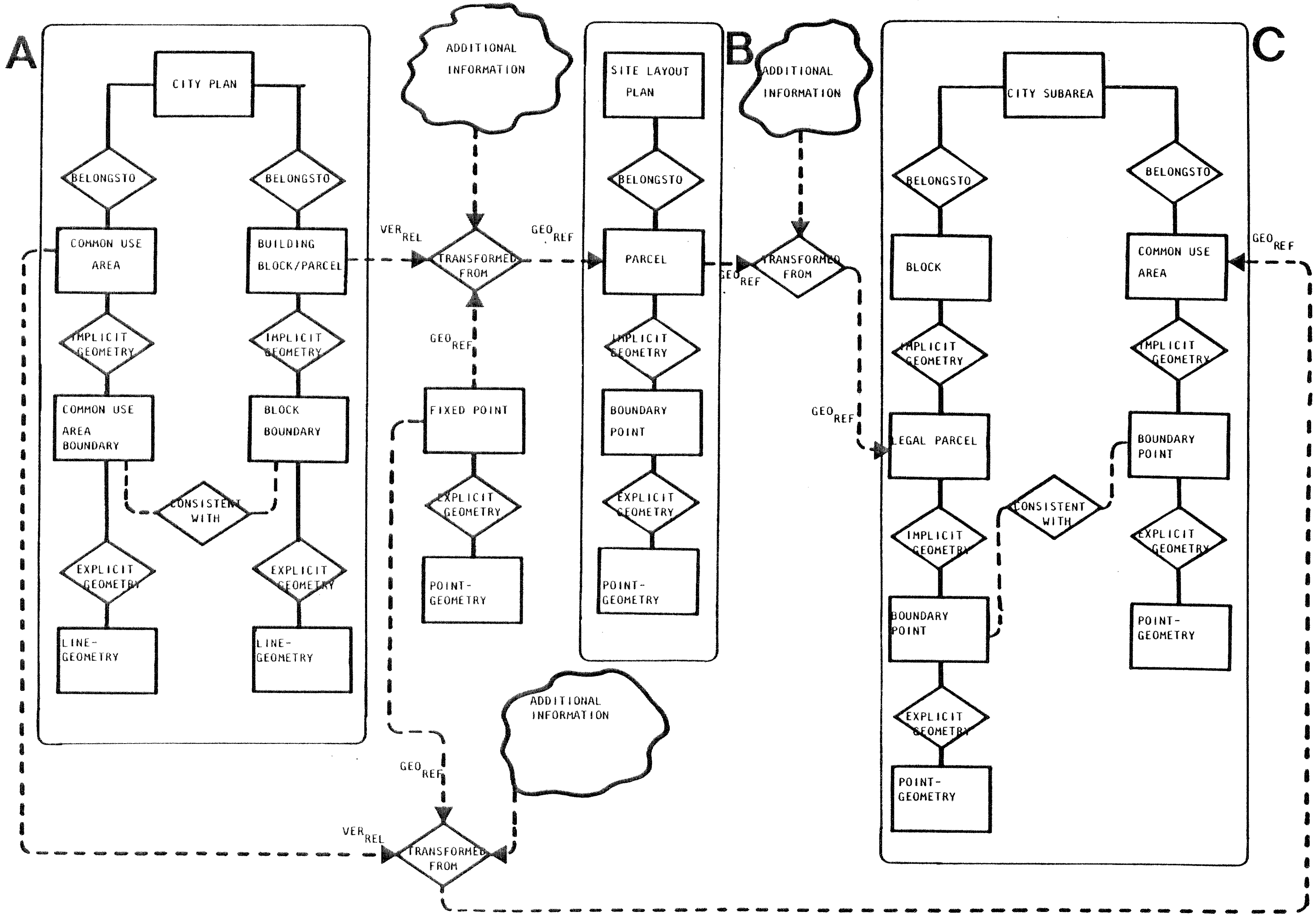

Fig. 5. Dynamic conceptual schema about land use planning.

These change processes differ from the previous example because of the interaction. Transformations from A to B and B to C can not be performed without human intervention and additional information. Because of the additional information also the data accuracy can be improved: draft coordinates have great tolerances (verification accuracy); site layout plans are computed by using fixed point information from the data base and the accuracy is reduced to geometric. The dominating accuracy type is geometric. In addition to these the accuracy is changed from relative to reference.

5. CONCLUSIONS

Geodata base dynamics is here understood as the evolution of entity representations. The two topics treated - data accuracy and structure conversions - make the "skeleton" of the dynamics.

In this study we have identified a certain taxonomy of data accuracy types. The following statements should be valid in any land information system:

- Information registration does not improve accuracy. Positioning and geometric accuracies can be high but the total accuracy can be dominated by interpretation and verification accuracies.
- Geometric accuracy can be specified. Positioning accuracy can be quantified exactly based on calibrations and system accuracy in some experimental range.
- Interpretation accuracy can be improved drastically by comparing the registered data with original objects.
- Verification accuracy is application dependent and the same information can be represented by varying verification accuracies depending on the generalization level, scale, use, etc.
- Verification accuracy can be changed during the processing (e.g. generalization, editing, drafting), it can be both decreased and increased especially in interactive process.
- In a closed process application accuracy can not be improved. E.g. information output process with high geometric accuracy possibilities does not effect on the interpretation errors or verification tolerances. Thus the plotter device and materials (pens, linewidths) have to be chosen appropriately.
- Relative, internal accuracy of a data set can be very high but the reference accuracy of the data in the question very low. The value of relative accuracy depends on the individual circumstances in one special data registration case. The selection data used as reference, to "globalize" the separate data sets must be based on thorough understanding of the application.

In this study we also gave a proposal for notations to be used in describing the dynamic data structures. On the results we can summarize the following:

- Data structure conversions are treated here as transformations between entity representations.
- Static data structures (and static conceptual schemas) are joined together with transformation relations.
- Transformations of entity representations can be based on interaction and information input but they can also be performed as closed processes.
- The dynamics of an integrated land information system and its geodata base must be analysed and described, otherwise a drastic amount of data redundancy and incompatability occurs between separate applications.
- Only a thorough semantic analysis of the object system can lead to a realistic dynamic model.

In this study we did not deal with accuracy measurements or technical implementations of data structure conversions. The further study should be oriented towards the standardization of accuracy concepts, measuring units and methods as well as the standardization of the logical data descriptions including both spatial and temporal dimensions.

7. REFERENCES

Chrisman, N., "A Theory of Cartographic Error and Its Measurement in Digital Data Bases", in Proceedings of AutoCarto 5, pp.159...168, 1982.

Chrisman, N., "The Role of Quality Information in The Long-Term Functioning of A Geographic Information System", in Proceedings of AutoCarto 6, pp. 303...311, 1983.

Deumlich,F., Surveying Instruments, de Gruyter, pp. 318, 1982.

Eloranta,K., "A Semantic View of on Geometric Representation of Urban Land Information", in Proceedings of Eurographics 82, pp.139...147, Manchester, 1982.

ISO/TC97SC6WG3: "Conceptual Schema in Data Base Management Systems", (edited by G.Sharman), draft (30.9.1980).

Nagy,G. & Wagle, S., "Geographical Data Processing" in Computing Surveys, Vol.11, no. 2, June 1979.

SFS 3700, Metrology. Vocabulary of fundamental terms of metrology. The Finnish Standardization Association, Helsinki, 1982, pp.225.

Wolfendale, P.C.F., "Machine Accuracies in Automatic Cartography", in Cartographic Journal, 1967:1, pp. 24...27.