KNOWLEDGE ACQUISITION IN MAP GENERALIZATION USING INTERACTIVE SYSTEMS AND MACHINE LEARNING

Tumasch Reichenbacher
Department of Geography
University of Zurich
Winterthurerstrasse 190
8057 Zurich (Switzerland)
FAX: +41-1-362 52 27
E-mail: tumasch@gis.geogr.unizh.ch

ABSTRACT

Current software for GIS and digital cartography suffers from a lack of intelligent generalization tools mainly due to the absence of formal generalization knowledge. Considering the increasing importance of digital spatial data at different scales, acquisition and formalization of generalization knowledge seems to be becoming a prerequisite for successful use of GIS and digital cartography in the future. This study presents a method for the acquisition of mainly procedural generalization knowledge in an interactive environment. The focus here is on techniques such as user interaction logging and machine learning. Regarding knowledge acquisition (KA) as a complex process, the proposed method integrates the different stages in a comprehensive framework. These stages are problem identification, interaction logging, editing of the recorded data, interpretation, implementation of knowledge and testing. This paper describes a simple experiment which demonstrates the feasibility of the method by considering the parameter choice for the Lang algorithm.

INTRODUCTION

In recent times mapping techniques no longer belong exclusively to cartographers. With the widespread use of Geographical Information Systems (GIS) and their possibilities, users are more and more becoming cartographer themselves. The problem is that most users have none, or very little knowledge of cartographic design. This might lead to unsatisfactory products or even to wrong decisions (Weibel and Buttenfield 1992). Current GIS do not offer sophisticated tools to support the inexperienced user in the task of map design. This is especially the case for the complex process of map generalization. McMaster and Shea (1988) define generalization in a digital environment as

> "the application of both spatial and attribute transformations in order to maintain clarity, with appropriate content, at a given scale, for a chosen map purpose and intended audience."

The need to assist the generalization process with the computer has been postulated by many researchers in this field. A first approach at tackling the problem of user support in map generalization was expert system technology. Apart from many attempts few systems have been successfully implemented. The main reason for the failure of expert systems for map generalization is probably the lack of explicit and formalized knowledge. Problems with acquisition and formalization of generalization knowledge are related to the characteristics of generalization: the process is ill-defined and represents a semi-structured problem (Armstrong 1991). To overcome some of the problems of a pure expert system approach, Weibel (1991) proposed a different concept, termed amplified intelligence. In contrast to expert systems, this concept is not a fully automatic one, but allows interactive generalization. Still, the crucial element is knowledge and the problem of the *knowledge acquisition bottleneck* has not been resolved yet. Therefore a lot of attention has been given to the formalization of generalization knowledge in the last few years.

In the realm of map generalization a few projects and studies have been conducted aiming to formalize generalization knowledge. A review of knowledge acquisition methods for map generalization can be found in Weibel (1993). Guidelines in verbal form and graphical form have been developed by several National Mapping Agencies. Buttenfield et al. (1991) and Leitner (1993) try to extract knowledge by comparing map series and built inventories for that purpose. The use of a graphical user interface for the acquisition of parameter values of simplification algorithms has been proposed by Chang and McMaster (1993).

In the knowledge acquisition literature also other methods are proposed (Welbank 1983; McGraw and Harbison-Briggs 1989; Boose 1989). Most of the techniques, like interviews for example, are not well suited to the cartographic domain. One technique however, process tracing and protocol analysis, can be adapted to the domain of map generalization in order to extract procedural knowledge. The methodology presented in the next section will discuss the use of that technique.

A METHODOLOGY FOR KNOWLEDGE ACQUISITION BY INTERACTIVE PROCESS TRACING

A methodology for cartographic knowledge acquisition was designed and implemented which integrates different, to date isolated methods into a comprehensive framework. The underlying idea is to log the interactions between an expert user and a generalization system during a working session and later interpret the logs using inductive learning algorithms. In solving the given generalization problem with the help of the operators available in the interactive system, the user is capable of contributing his/her subjective (procedural) knowledge about the generalization process (Figure 1). Shape measures are used to achieve a description of the structure of the features in the cartographic database (structural knowledge). Thus, the procedural knowledge can be traced and related to structural knowledge, allowing one to answer questions, such as which generalization operators are selected and which parameter values are specified for a particular operator in relation to scale, feature class, and line complexity.

The proposed methodology is based on the extension of two frameworks. It basically represents an extension to the general model of knowledge acquisition by Buchanan et al. (1983). For the matters of generalization the Brassel and Weibel framework (Brassel and Weibel 1988) is used as a model of the generalization process. The schema of the developed methodology is shown in Figure 1.

The first stage in the knowledge acquisition process is the **problem identification**. This includes a knowledge level analysis of the generalization process to reveal the different types and levels of knowledge and tasks. The goal is to identify adequate knowledge acquisition techniques that can be applied. Task level frameworks developed by McMaster and Shea (1992), Beard (1991) and João (1990) offer a decomposition of the overall generalization task in generalization operators, generalization algorithms and algorithm parameters allowing a further structuring of the problem domain. (Armstrong 1991) and (Muller 1991) identify three basic knowledge types in cartography: geometrical, structural and procedural knowledge. These knowledge types can be linked to the Brassel and Weibel model.

The analysis of the generalization process in the problem identification step indicates possible means for knowledge acquisition. The technique for the acquisition of procedural knowledge used is **process tracing**. However, in contrast to traditional knowledge acquisition, where process tracing involves the recording of verbal data (McGraw and Harbison-Briggs 1989), in this context interactions between user and generalization system will be logged. Thus, this method forms a special instance of process tracing in traditional knowledge acquisition. Therefore a technique and data format have been designed, allowing the recording of interactions such as generalization operator selection and parameter settings, as well as descriptive measures characterizing the structure of the features processed. The logged data can be traced to identify the decisions and tap the procedural knowledge governing this process.

Editing and preprocessing encompass tasks such as format conversion, noise removal, and data classification in order to prepare the logged data for the subsequent interpretation step. This stage is quite crucial for the overall efficiency of the methodology as it can be very tedious. A powerful user interface should support the user in this task.

The **interpretation** of the previously logged data makes use of inductive machine learning algorithms to gain more structured knowledge. Considering the fact that the number of interactions logged in a session quickly reaches several hundreds, an automated approach is certainly justified. The underlying concept of the algorithms is *learning by example* The logged interactions between user and system form examples of procedural generalization knowledge. The algorithms try to find similarities in the given examples and induce a more general description in the form of decision trees or production rules. The advantage of inductive tools are that they have been studied and tested for a long time in various projects in machine learning and artificial intelligence, and that they produce easily interpretable output. For this methodology implementation of algorithms available on the Internet have been considered.

In **knowledge implementation**, the prototype rules which had been derived previously are implemented within the interactive system (or possibly a different system). The knowledge has not necessarily to be represented as production rules. Possible representation formalisms are *logic, procedural methods, object-oriented methods, semantic nets, frames*. Shea (1991) presents a *parameter table structure* for the representation of declarative and procedural generalization knowledge. The acquired knowledge could also be incorporated into new generalization algorithms rather than building a knowledge base.

Finally, **knowledge testing and evaluation** assess the quality of the derived knowledge, attempting to identify inconsistencies and rule conflicts or even missing knowledge. This step may also include an evaluation of the system performance. The evaluation of the knowledge is strongly related to questions about the quality of maps and generalization solutions. Implemented knowledge must be critically reviewed by cartographic experts.

Obviously, the KA methodology does not represent a strict sequence of steps; various possibilities for feedback loops exist (Figure 1). The difference of this methodology compared to other KA approaches in the generalization domain is that the human expert is kept in the loop. Thus, the expert can always provide further comments and guidance (e.g. declaration of specific structural attributes).

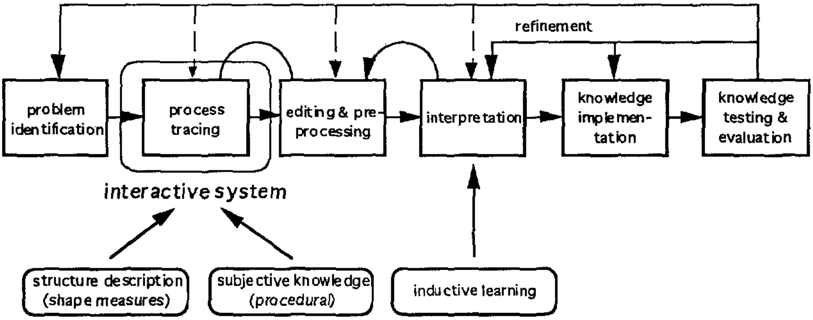

refinement
problem identification
process tracing
editing & pre-processing
interpretation
knowledge implemen-tation
knowledge testing & evaluation
interactive system
structure description (shape measures)
subjective knowledge (procedural)
inductive learning

Figure 1: A methodology for knowledge acquisition and formalization in cartographic generalization

IMPLEMENTATION

The methodology was implemented as a module of an interactive generalization system developed in a previous project by Schlegel and Weibel (1995). This system was designed according to the concept of amplified intelligence (Weibel 1991) and uses a commercial GIS as a carrier system.

The knowledge acquisition module (Figure 2) is based on an interface to control the interactive session using GUI elements (menus, check boxes, sliders, etc.). The main menu of the interface (Figure 2, top left) allows the user to invoke a set of dialogs to define the global control parameters, calculate and display structure measures (Figure 2, top right), browse through feature attribute tables, perform generalization (Figure 2, left) and log his/her interactions. Figure 2 shows also the display of the calculated structure measures (center window). In the background a line data set is drawn and on the bottom right window a zoomed extraction thereof.

Based on problem identification, an adequate logging format was implemented to enable logging the interactions of an user with the system. This involves the definition of an interaction, the identification of relevant items to log and a mechanism to log these data during an interactive session. An *interaction* is any operation the user performs in the system as response to or to invoke a system action. Examples of interactions include the selection of features, choosing a generalization operator and algorithm, specifying parameters values for an algorithm, or defining feature symbology. Relevant information is interaction and structural description of the map features. Each performed interaction is written along with further describing attributes to an *interaction table* (Table 1). For each map feature structural description is stored in attributes added to the *feature attribute table* of the GIS database system. Global generalization control information such as map scale and map purpose are stored in a *map table.* The structure measures are an implementation of Buttenfields structure signatures (Buttenfield 1991) and measures proposed by McMaster(1986). The knowledge acquisition interface also faciliates the contribution of subjective knowledge by the user. The user can define his own attributes to describe the structure of map features. Furthermore a possibility to attach comments to special or difficult cases exists by linking a text file to a map feature.

Interaction Table	**Feature Attribute Table**	**Map Table**
(stored for each **interaction**)	(stored for each **feature**)	(stored for each **map**)
interaction type (algorithm)	feature-id	map_id
user-id	map-id	map_purpose
map-id	**length** (line length)	source_scale
feature-id	**coor** (number of coordinates)	target_scale
date	**anchor** (anchor line length)	
time	**seg** (segmentation)	
parameters	**band** (bandwidth)	
	ev (error variance)	
	cv_ev (coefficient of variance)	
	lratio (length ratio)	
	asl (average segment length)	
	cv_asl (coefficient of variance)	
	aa (average angularity)	

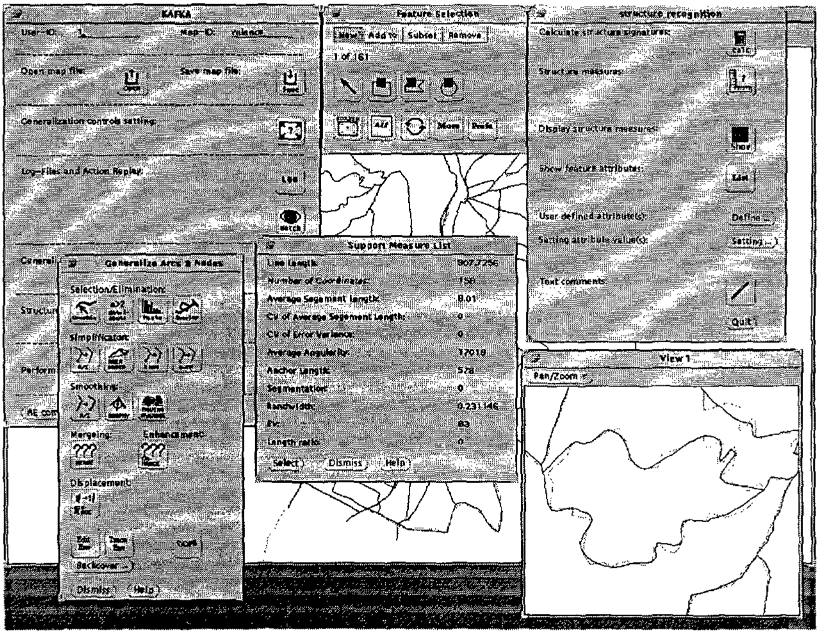
Table 1: Data logged for interactions, features and maps

EXPERIMENT

An experimental study with the following objectives was carried out: 1) empirically assess the feasibility of the proposed methodology, 2) extract prototype rules, and 3) obtain experience in the application of the techniques involved.

Test data from the French National Mapping Agency (Institut Géographique National – IGN) were used for the experiment. Two data sets were available for the study area located in the region of Valence in the Rhone valley. The first data set represents an extraction from IGN's BD Carto product at a scale of 1: 100,000. To keep the experiment at a manageable level of complexity, only the road network was selected. The second data set contains the road network for the same area which was manually generalized to a scale of 1:250,000 and later digitized. In this experiment the manual result served as the target generalization (see Figure 2, bottom right window).

Since the emphasis of this initial study was on technical feasibility assessment rather than knowledge acquisition, the experimental setup was simple. A number of constraining assumptions were made. Firstly, the source data set is presumed to be cartographically correct. As the data is from a NMA this requirement is certainly fullfilled. Secondly, it must be possible to perform a generalization using the simplification and smoothing algorithms available in the prototype system (Schlegel and Weibel 1995). This is necessary to trace back the decisions. Thirdly, as mentioned above, only a single feature class (road network) is involved. Finally, it is assumed that the user generalizes the source data using the manual result as a template (backdrop); the task is to try to match this solution applying the tools available in the system. This latter constraint was introduced to simplify the KA process, but would need to be relaxed in future experiments.



Support Measure List

Line Length:	907.7256
Number of Coordinates:	158
Average Segement Length:	8.01
CV of Average Segement Length:	0
CV of Error Variance:	0
Average Angularity:	17018
Anchor Length:	578
Segmentation:	0
Bandwidth:	0.231146
Ev:	83
Length ratio:	0

Figure 2: The knowledge acquisition interface

Problem identification

On the task level we can identify one feature class (roads), one operator (simplification), one algorithm (Lang) and two parameters. Initially, only the Lang algorithm (Lang 1969) was chosen since it uses two parameters (instead of one like most other algorithms) — a distance tolerance and a look-ahead number — and preserves the character of cartographic lines well (McMaster 1987). The problem to solve is setting the two parameters of the Lang algorithm to control the simplification of the lines representing a road network. The goal is to match as best as possible the target generalization.

Process tracing

The source data set contains 161 cartographic lines. 52 lines were generalized using the interactive system. For each line, parameters were selected using the GUI sliders so as to best approximate the manual generalization displayed as a visual backdrop. Since the interactive simplification of a line usually represents a trial-and-error process, with several tries rejected until the optimal parameter setting is found, an 'undo' mechanism was implemented. For this purpose an 'interaction cycle' was defined to include all interactions between the selection of a feature to its deselection. Interactions rejected by the undo operator are tagged as invalid in the corresponding record of the interaction table and removed in the subsequent preprocessing step.

Editing and preprocessing

Before the logged data can be passed to inductive learning algorithms, further data editing and reformatting is necessary. First, the interaction table is parsed and the invalid interactions are deleted. The result is a table containing one record for each line (52 in our case). Using the `feature_id` as a key, the *interaction table* is then joined with the *feature attribute table* in order to also access the structural information calculated before (see Table 1). In our experiment, the following attributes were subsequently written to a text file (Table 2).

Since only one algorithm was used for a single map data set, the attributes `interaction`, `map_purpose`, `source_scale` and `target_scale` were not included as they remain constant. The inductive learning algorithms used in this experiment require input in the form of a list of symbolic descriptions of an example. The structural data, however, were measured on a metric scale. The scale of measurement was therefore transformed by a k-means cluster analysis, resulting in two or three classes per attribute with symbolic labels such as high, low, medium, few etc. The file was then formatted to the final list structure in a text processor.

t	n	coor	length	anchor	lratio	band
seg	ev	cv_ev	asl	cv_asl	aa	road_class

Table 2: Attributes for the interpretation with machine learning

Interpretation

Interpretation was performed using public domain inductive learning tools. These tools running under Macintosh Common LISP represent implementations of ID3 (Quinlan 1986), AQ15 (Hong et al. 1986), and PRISM (Cendrowska 1987), which have the advantage of producing easy-to-read production rules. The goal of this step is to organize the knowledge contained in the unstructured interactions.

Our experiments initially focused on deriving prototype rules for the selection of the distance parameter `t` of the Lang algorithm. The conditional part of the rules then contains structural attributes. Two examples out of the 10 rules generated by the AQ15 algorithm are shown here.

Rule 1:

```
(IF
  ((COOR FEW) and (LRATIO MEDIUM) and (CV_EV LOW)) or
  ((ASL LONG) and (CV_ASL HIGH) and (CV_EV MEDIUM))
 THEN (CLASS T8))
```

Rule 2:

```
(IF
   ((LRATIO BIG)) or
   ((COOR MEDIUM) and (ANCHOR MEDIUM) and (CV_EV MEDIUM)) or
   ((SEG MEDIUM) and (EV LOW) and (LRATIO MEDIUM))
 THEN (CLASS T16))
```

The target CLASS into which the algorithm tries to classify the data represents the value for the distance tolerance of the Lang algorithm (denoted by T#). This notation is necessary, because the inductive learning algorithms only process symbolic values. However, the number after T is the distance value in map units (1MU = 10 m), e.g. 160 meters in Rule 2. Rule 1 states that the distance parameter is equal to class 8 if the line has only few coordinates, the length ratio (the ratio of the line length and the anchor line length) is medium, and the coefficient of variation of the error variance is low. Furthermore, the same distance tolerance applies if the average segment length is long, the coefficient of variation of the average segment length is high, and the coefficient of variation of the error variance is medium. Rule 2 can be read analogously. It should also be noted that the symbolic values (high, medium, long, etc.) which may seem somewhat fuzzy could easily be constrained by numeric upper and lower bounds of the corresponding class and range checked by the inference engine of the knowledge-based system which serves for implementation of the prototype rules. Likewise, the number of classes for each attribute could be increased to achieve greater resolution.

Knowledge implementation & knowledge testing and evaluation

The two final steps of the proposed methodology — knowledge implementation, and knowledge testing and evaluation — were not carried out yet in this experiment. Thus, the prototype rules have been neither verified nor falsified. It should be noted, however, that the prototype knowledge extracted by inductive learning does not necessarily need to be implemented in the form of production rules. In the context of systems such as the experimental platform used here, an algorithmic implementation via adaptive defaults retrieved from a lookup table according to the structural measures of cartographic features may be more appropriate.

SUMMARY AND CONCLUSIONS

While the prototype rules generated in this simple experiment are certainly not yet very sophisticated, the experiment allows to clearly see the technical feasibility of this novel approach. In order to be useful to knowledge acquisition, further research must be carried out with the dual objective of achieving technical improvement as well as running more sophisticated KA experiments.

In technical terms, the major critical factors of the methodology are the interactive generalization system and the methods for achieving the structural descriptions of map features. The generalization system used here is still rather simple. It would need to be extended by further generalization operators based on more powerful data models. Additionally, more research on user interfaces for generalization is required. With respect to structural feature description, better shape measures must be developed and implemented. For instance, the measures by Buttenfield (1991) used in this study are severely biased by the fact that they are based on the 'anchor line' (the line connecting the endnodes of a line), which represents a poor approximation of the general trend of the cartographic line. The measures developed by Plazanet (1995) seem much more promising. Finally, structure description should also be enriched by contextual information (e.g., topology, neighboring features, feature clustering, etc.). The modelling of semantics of features is again strongly related to more powerful data models.

In terms of KA experiments, future research needs to relax the above constraining assumptions and increase the range of generalization operators and feature classes under study. Furthermore not only single interactions should be taken into account, but also the sequencing of them. Great care must be taken to design meaningful experiments in collaboration with expert cartographers. Additionally, future KA studies must involve significant numbers of cartographers. Under such terms, it will eventually be feasible to address the issues of knowledge implementation as well as testing and evaluation. Evaluation methods from software and knowledge engineering should be considered as well as new evaluation methods for quality assessment in the generalization process. Instead of developing new inductive learning algorithms it seems more appropriate to evaluate in further studies the most suitable and efficient algorithms available.

ACKNOWLEDGEMENTS

The data used in the experiment was kindly provided by the research staff of the COGIT Labs at the Institut Géographique National, Saint-Mandé (France). I would also like to thank Robert Weibel for encouraging me to write this paper.

REFERENCES

Armstrong, M.P. (1991). Knowledge classification and organization. In: Buttenfield, B. P., McMaster, R.B., (Eds.) *Map Generalization - Making Rules for Knowledge Representation*. Longman. 86-102.

Beard, K. (1991). User Interaction in Map Generalization. *Proceedings of the 15th Conference of the International Cartographic Association*, Bournemouth.

Boose, J.H. (1989). A survey of knowledge acquisition techniques and tools. *Knowledge Acquisition*, **1** : 3-37.

Brassel, K.E., Weibel, R. (1988). A review and conceptual framework of automated map generalization. *International Journal of Geographical Information Systems*, **2** (3): 229-244.

Buttenfield, B.P., Weber, C.R.,Leitner, M.,Phelan, J.J., Rasmussen, D.M., Wright, G.R. (1991). How Does a Cartographic Object Behave? Computer Inventory of Topographic Maps. *Proceedings of the GIS/LIS '91*, Atlanta, 2: 891-900.

Buttenfield, B.P. (1991). A rule for describing line feature geometry. In: Buttenfield, B. P., McMaster, R.B., (Eds.) *Map Generalization - Making Rules for Knowledge Representation*. Longman. 150-171.

Chang, H., McMaster, R.B. (1993). Interface Design and Knowledge Acquisition for Cartographic Generalization. *Proceedings of the AutoCarto 11*, Minneapolis, 187-196.

João , E.M. (1990). What expert systems don't know: the role of the user in GIS generalisation. *Proceedings of the NATO ASI on Cognitive and Linguistic Aspects of Geographic Space*, Las Navas del Marques, 493-506.

Lang, T. (1969). Rules for robot draughtsmen. *Geographical Magazine*, **62** (1): 50-51.

Leitner, M. (1993). *Prototype Rules for Automated Map Generalization*. Master's Thesis, NCGIA, State University of New York, Buffalo, NY.

McGraw, K.L., Harbison-Briggs K. (1989). *Knowledge Acquisition- Principles and Guidelines*. Prentice - Hall International.

McMaster, R.B. (1986). A statistical analysis of mathematical measures for linear simplification. *The American Cartographer*, **13** (2): 103-116.

McMaster, R.B. (1987). The Geometric Properties of Numerical Generalization. *Geographical Analysis*, **19** (4): 330-346.

McMaster, R.B., Shea, K.S. (1988). Cartographic Generalization in a Digital Environment: A Framework for Implementation in a Geographic Information System. *Proceedings of the GIS/LIS '88*, San Antonio (TX), 240-249.

McMaster, R.B., Shea, K.S. (1992). *Generalization in Digital Cartography*. Association of American Geographers.

Muller, J.-C. (1991). Generalization of Spatial Databases. In: Maguire, D. J., Mark, D., Goodchild, M.F., (Eds.) *Geographical Information Systems: Principles and Applications*. London, Longman. 269-297.

Schlegel, A., Weibel, R. (1995). Extending a General-Purpose GIS for Computer-Assisted Generalization. *Proceedings of the 17th International Cartographic Congress of the ICA*, Barcelona.

Shea, S.K. (1991). Design considerations for an artificially intelligent system. In: Buttenfield, B. P., McMaster, R.B., (Eds.) *Map Generalization - Making Rules for Knowledge Representation*. Longman. 3-20.

Weibel, R. (1991). Amplified Intelligence and Rule-Based Systems. In: Buttenfield, B. P., McMaster, R.B., (Eds.) *Map Generalization - Making Rules for Knowledge Representation*. Longman. 172-186.

Weibel, R., Buttefield, B.P. (1992). Improvement of GIS graphics for analysis and decision-making. *International Journal of Geographical Information Systems*, **6** (3): 223-245.

Weibel, R. (1993). *Knowledge Acquisition for Map Generalization: Methods and Prospects*. NCGIA Research Initiative 8. NCGIA SUNY Buffalo (NY).

Welbank, M. (1983). *British Telecom Report on Knowledge Acquisition*. 1983. British Telecom, London.