Classification:

Assumptions and Implications for Conceptual Modeling.

by

Tor Kristian Bjelland

Dissertation in Information Science

Department of Information Science and Media Studies
Faculty of Social Science
University of Bergen

2004

Submitted to the Department of Information Science and Media Studies, University of Bergen, on October 8, 2004
in Partial Fulfillment of the Requirements for the Degree of Dr. Polit.

Supervised by:

Joan C. Nordbotten
Department of Information Science and Media Studies
Faculty of Social Science, University of Bergen

Abstract.

Classification is generally held to be of fundamental importance to the analysis and design of software applications. This is clearly reflected in data modelling terminology, in which terms like *classification*, *concept*, *class*, *superclass*, *subclass*, *IS_A relationship*, *generalization* and *specialization* are frequently used. But what exactly does classification mean?

This thesis contains three studies, all of which are based on the assumption that classification has not received sufficient attention by the data modelling community.

The first study analyzes the concept of classification from the perspectives of diverse disciplines, such as psychology, linguistics, archaeology, artificial intelligence and data bases. The study uncovers four different senses of classification, and leads to the analyses of related concepts, such as *concept*, *class*, *type*, *object* and *property*. The implications that follow from the study suggest that classification may:

1. Contribute to a shared understanding of basic modelling concepts
2. Result in a vocabulary that may be formally verified with respect to its completeness, logical consistence, and understandability.
3. Provide a basis for modelling decisions.
4. Emphasize socio-technical consequences and measures.
5. Enhance data integrity.
6. Support the validation and interpretation of conceptual models.
7. Support schema integration.

It is concluded that classification may be viewed as a prerequisite to conceptual modelling, and that the conceptual modelling process should be divided into two, separate tasks: the classification task, which is concerned with the definitional properties of concepts, and the modelling task, which, is concerned with the descriptive properties of the objects to which the concepts applies.

The second study is conducted to test the initial assumption that classification is not properly attended to by the data modelling community. A content analysis of 29 text books on conceptual modelling and database design reveals that none of the text books contain explicit definitions of classification in all of the four senses identified in the first study.

Based on the findings from the first two studies, a methodology that integrates classification and conceptual modelling is developed and presented in a separate chapter. The chapter provides a theoretical justification for a constructivist perspective on classification and conceptual modelling, explains its theoretical concepts, and describes the method through a set of guidelines and examples. The examples address the first five implications and demonstrate the pragmatic utility of the conceptual framework developed in the first study.

The third study is conducted to empirically test the sixth implication, i.e., the effect of classification on interpretation tasks. It is shown that people who know the membership conditions will make other judgments than people for which the membership conditions are unknown. It is also shown that without knowledge of membership conditions people become less confident, and less consistent in their interpretations, exhibiting larger variation in their judgments.

Table of contents

Abstract iii

Table of contents iv

List of figures vii

List of tables viii

Acknowledgements ix

1.0 Introduction 1

1.1 Background 2

1.2 Related work 5

1.3 Assumptions and motivations 6

1.4 Research questions and methods 7

2.0 Concept analysis of classification 9

2.1 Purpose 10

2.2 Method 11

2.2.1 Identification of the concept of interest and associated expressions 11

2.2.2 Selection of setting and sample 15

2.2.3 Data Collection 16

2.2.4 Analysis 18

2.3 Findings 19

2.3.1 Metaphysical perspectives on classification 19

2.3.2 A framework for analysis of classification 21

2.3.2.1 Cognitive perspectives on classification 21

2.3.2.2 Practical and theoretical perspectives on classification 28

2.3.2.3 Representational perspectives on classification 39

2.4 Discussion 48

2.4.1 Definitions 48

2.4.2 The resulting conceptual framework 55

2.5 Implications 56

2.6 Conclusion 61

3.0 Classification concepts in conceptual modeling 62

3.1 Introduction 62

3.2 Method 62

3.2.1 Identification of text books 62

3.2.2 Analysis 63

3.3 Findings 64

3.3.1 Classification – a process, a product, a judgment, and a relation 65

3.3.2 Constructs – concept, class, object, and property 71

3.3.3 Classification versus conceptual modeling 81

3.4 Overall summary and conclusions 84

4.0 Including classification concepts in conceptual modelling 86
4.1 Introduction 86
4.2 Positivist versus constructivist perspectives on conceptual modelling 86
4.2.1 What is the form and nature of the reality which conceptual models are supposed to model? 89
4.2.2 What is the relationship between the designer and the users' mental models? 90
4.2.3 How is the conceptual modelling process conducted? 91
4.2.4 Conclusion 92
4.3 Conceptual modelling – theoretical aspects 93
4.3.1 Basic concepts associated with classification 95
4.3.2 Classification versus modelling 100
4.3.3 Benefits of classification 101
4.4 Conceptual modelling – practical aspects 103
4.4.1 Collect user requirements 104
4.4.2 Identify concepts 105
4.4.3 Define the concepts 108
4.4.4 Evaluate the vocabulary 112
4.4.5 Specify types and complete the conceptual model 113
4.4.6 Validate the model according to the vocabulary 116

5.0 Empirical study of interpretation tasks 117
5.1 Introduction 117
5.2 Related work 117
5.3 Research question 118
5.4 Experimental design and framework 120
5.4.1 Independent variable 121
5.4.2 Dependent variables 121
5.4.3 Task characteristics 122
5.4.4 Subjects 122
5.4.5 Hypotheses 123
5.4.6 Experimental results 125
5.5 Overall interpretation of the results 139
5.6 Conclusion 140

6.0 Overall discussion 141
6.1 Introduction 141
6.2 Implications for theory and practice 142
6.2.1 Shared understanding of basic concepts 142
6.2.2 Complete, logical consistent and understandable vocabulary 142
6.2.3 Basis for modelling decisions 143
6.2.4 Socio-technical consequences and measures 143
6.2.5 Data integrity 143
6.2.6 Validation and interpretation of conceptual models 144
6.2.7 Schema integration 144
6.3 Validity and reliability 145
6.3.1 Concept analysis of classification 145
6.3.2 Content analysis of text books on conceptual modelling 147
6.3.3 Effects of classification on interpretation tasks 148
6.4 Recommendations and future research 150

Appendices 151

Appendix A: Instrument to measure of interpretation tasks 151
Appendix B: Lecture held between experiment 1 and experiment 2 175
Appendix C: SPSS Output files from experiment 1 and 2. 199
Appendix D: Instruments to measure design tasks 203
Appendix E: Terminology database to support classification 217
Appendix F: Alternative diagrammatic notations 223

References 226

List of figures

Figure 2.1: Example data collection sheet 17
Figure 2.2: Analysis framework for classification and related notions 21
Figure 2.3: Classification and related concepts from the perspective of cognitive science 27
Figure 2.4: Classification and related concepts from the perspective of terminology 31
Figure 2.5: Classification and related concepts from the perspective of archaeology 37
Figure 2.6: Classification and related concepts from the perspective of intuitive classification 46
Figure 2.7: Classification and related concepts from a constructivist perspective 46
Figure 2.8: Example of a binary relationship 59
Figure 2.9: Example binary relationship with membership conditions included for entity types 59

Figure 4.1: The conceptual data modelling process I 91
Figure 4.2: The conceptual data modelling process II 94
Figure 4.3: A conception, its intension and extension 95
Figure 4.4: A concept and a conception, its intension and extension 95
Figure 4.5: A modified conception, based on concept, its intension and extension 96
Figure 4.6: The form of a terminological entry 96
Figure 4.7: Examples of intensional and extensional definitions 97
Figure 4.8: Classification as a system of labelled concepts 98
Figure 4.9: Classification as a system of indented concepts 98
Figure 4.10: Classification versus generalization 99
Figure 4.11: Classification versus modelling 100
Figure 4.12: Extended type construct 101
Figure 4.13: The baseball model. 103
Figure 4.14: Fact-finding techniques 104
Figure 4.15: An order form containing primitive concepts arranged in clusters 107
Figure 4.16: A simple definition and an associated concept system 109
Figure 4.17: Example form for the definition and extra information about a term 111
Figure 4.18: Definition of terms denoting relationships 112
Figure 4.19: Example concept system 115
Figure 4.20: Example conceptual model 115

Figure 5.1: Informative and non-informative membership conditions 119
Figure 5.2: Experimental framework 120
Figure 5.3: Number of benchmark choices reported by the treatment group and control group 126
Figure 5:4: Number of benchmark choices for informative interpretation tasks 127
Figure 5.5: Number of benchmark choices for non-informative interpretation tasks 127
Figure 5.6: Reported confidence level for each choice 128
Figure 5.7: Reported confidence level for each choice based on informative conditions 129
Figure 5.8: Reported confidence level for each choice based on informative conditions 129
Figure 5.9: Number of benchmark choices reported by the treatment group and control group 134
Figure 5:10: Number of benchmark choices for interpretation tasks with informative membership conditions 135
Figure 5.11: Number of benchmark choices for interpretation tasks with non-informative membership conditions 135
Figure 5.12: Reported confidence level for each choice 136
Figure 5.13: Reported confidence level for each choice based on informative membership conditions 136
Figure 5.14: Reported confidence level for each choice based on informative membership conditions 136

List of tables

Table 2.1: Main search terms 11
Table 2.2: Surrogate table for construct terms 12
Table 2.3: Surrogate table for process terms 12
Table 2.4: Gross distribution of articles sorted by content 15
Table 2.5: Distribution of articles sorted by Journal 16
Table 2.6: Key concepts in psychological theories of classification 25
Table 2.7: Key concepts in terminology 32
Table 2.8: Key concepts in archaeology related to classification structures and arrangements 34
Table 2.9: Key concepts in archaeology related to classification processes. 34
Table 2.10: Representational, mental, and physical elements related to the notion of typehood 35
Table 2.11: Key concepts in archaeology related to the concretization of types (concepts) 35
Table 2.12: Key concepts in archaeology related to mental concepts 36
Table 2.13: Concepts in archaeology to denote objects 36
Table 2.14: Key concepts related to object-oriented analysis and design 44
Table 2.15: Key concepts in AI related to knowledge representation 45
Table 2.16: General ideas that are commonly related to classification 48
Table 2.17: A tentative framework to analyze schema integration problems 60

Table 3.1: List of authors 62
Table 3.2: General ideas of classification and associated indicator terms 63
Table 3.3: Findings and coding information 64
Table 3.4: Frequency distribution of positive scores on the four senses of classification 65
Table 3.5: Basic concepts and interpretations needed to conceive classification 71

Table 4.1: Schlaer and Mellor's list to identify classes and objects 105
Table 4.2: Ross' list to identify classes and objects 105
Table 4.3: Coad and Yourdon's list to identify classes and objects 106

Table 5.1: Number of subjects for each treatment in experiment 1 122
Table 5.2: Number of subjects for each treatment in experiment 2 122
Table 5.3: Subjects' self-reported expertise. 123
Table 5.4: Frequency table over the responses from the control group on experiment 1 125
Table 5.5: Frequency table over the responses from the treatment group on experiment 1 126
Table 5.6: Descriptive statistics from experiment 1 130
Table 5.7: Results from the t-tests on experiment 1 131
Table 5.8: Frequency table of the control group responses in experiment2 133
Table 5.9: Frequency table of the treatment group responses in experiment2. 134
Table 5.10: Descriptive statistics from experiment 2 138
Table 5.11: Results from the t-tests on experiment 2 138

Table 6.1: Criteria for assessing the quality and rigour of research 145

Acknowledgements

I would like to express my gratitude to my advisor Joan Castro Nordbotten for giving me the opportunity to do my doctoral thesis at the Institute for Information and Media Science at the University of Bergen. Her initiative, endurance, guidance, and helpful discussions throughout the last seven years are highly appreciated.

I am also very grateful to Conrad Morgan for reviewing my paper to the WSES conference and for nominating me to the ICIS Doctoral Symposium in 2001. Many thanks also to Maung Sein for prompt and helpful advice on my paper to the Consortium.

Also, special thanks to Mike Spector for his eminent feedback on my philosophy of science essay, and to Pål Davidsen and Ragnar Fjelland for their contributions during the philosophy of science seminar.

I would also like to thank Cheryl L. Dunn for her time and effort to discuss research design and interpretation tasks with me, Tone Lønning for letting me use her lecturing hours to run my experiments, and Inge Thorsen for reviewing chapter 5.

Lastly, I am deeply indebted to Peter Larsen for reading my thesis as a whole, and for giving me valuable feedback for the final polish.

This work could not have been accomplished without financial and technical support from the University of Bergen and from Stord /Haugesund College. The support is highly appreciated.

Lastly, many thanks go to my family, Liv, Hanne and Torbjørn for their endurance and support, and to all of my colleagues for their ever lasting optimism on my behalf.

1.0 Introduction

Classification is generally recognized as a fundamental abstraction mechanism for conceptual modelling, and software engineering. (Booch, 1991; Mylopoulos, 1998). This is clearly reflected in data modelling terminology, in which terms like classification, concept, class, superclass, subclass, IS_A relationship, generalization and specialization are frequently used. Classification is also considered to be the hardest part of analysis and design (Booch, 1991). Yet, in spite of its importance, the discipline seems to lack a unified account of classification. As a result, the discipline is unable to provide simple answers to what classification means, how objects and classes are identified, and how class structures should be arranged and evaluated. As an example, consider the quote from Grady Booch, one of the leading figures among object-oriented methodologists:

> "Classification is the means whereby we order knowledge. In object-oriented design, recognizing the sameness among things allows us to expose the commonality within key abstractions and mechanisms, and eventually leads us to smaller and simpler architectures. Unfortunately, there is no golden path to classification. To the readers accustomed to finding cookbook answers, we unequivocally state that there are no simple recipes for identifying classes and objects. There is no such thing as a perfect class structure, nor the right set of objects...
>
> At a conference on software engineering, several developers were asked what rules they applied to identify classes and objects. Stroustrop, the designer of C++, responded: "It's a Holy Grail. There is no panacea". Gabriel, one of the designers of CLOS, stated, "That's a fundamental question for which there are no easy answer. I try things"." (Booch, 1991, p. 132).

From the quote above, one gets the impression that conceptual modelling, as a discipline, is in need of a unified vocabulary where terms like classification and related notions including concept, class, object and property are properly accounted for. To be able to rigorously reason about model constructs, to provide answers to questions about modeling, and justifications for claims, such as the ones cited above, it becomes necessary to specify the domain of discourse, in a logically consistent and coherent manner, (Sutcliffe 1994). The need for a unified vocabulary has been articulated by several authors, as expressed in the following quote:

> "Snyder notes that; "…the groups involved with OO lack a shared understanding of the basic concepts and a common vocabulary for discussing them". Yourdon warns that: "…there is still enormous variation (and some contradictions as well) between the notation, strategies, and semantics of the various OOAD methodologies"… Discussing inheritance, Winkler notes that: "…this key-concept of OOP is interpreted quite differently by different groups of the software community"… Ling and Teo also recognize the lack of standards as one of the main inadequacies in OO data models." (van Hillegersberg and Kumar, 1999, p. 113)

The motivation for this research is to study classification, and its implications with respect to conceptual modelling. Classification in this respect is understood as the process of defining the key concepts in an application domain. It is assumed that classification is a prerequisite to the data modelling process, and that the vocabulary, which results from classification will provide valuable help to designers engaged in the design of the conceptual data model. It is also assumed that the vocabulary associated with a conceptual data model will be of help to reviewers such as end users, or internal auditors in interpreting and validating existing conceptual data models.

The initial research approach is based on concept analysis of existing classification theories from the Cognitive sciences, Philosophy, Terminology, and Archaeology. The theoretical basis for the analysis is collected from reviews of approximately 50 text books and 250 papers from scientific journals and conference proceedings. Based on hypotheses derived from the concept analysis, an experimental research design is used to empirically test the effect of classification on interpretation tasks.

1.1 Background.

From a review of classification theories, it seems reasonable to distinguish between a cognitive, and a logical sense of classification. In the cognitive sense, classification is concerned with how people conceptualise the world, in the form of mental representations and operations. In the logical sense, classification is concerned with the definition of terms in order to concretise concepts. The main difference is that in the cognitive sense, concepts are subjective and private, while in the logical sense concepts are public, and hence, made inter-subjectively available by intensional definitions.

It appears that classification in the cognitive sense is the justification for classification in the logical sense. Research within the cognitive sciences has repeatedly demonstrated that concepts in general are subjective and vague, and liable to change, both between individuals and, over time, within the same individual. It is exactly this vagueness, instability, and subjectivity of mental concepts that cognitive theories of classification attempt to explain, and the logical theory attempts to overcome.

How does this relate to conceptual data modelling? First of all, the two senses of classification may be viewed as the starting and the end points of the conceptual data modeling process. Kroenke (1998) speaks of a database as a model of the user's mental models. Schlaer and Mellor (1988) view conceptual modeling as a process in which separate and sometimes conflicting conceptual frameworks are brought together. Hirschheim and Klein (1995), describe conceptual modeling as the fusion of horizons of meaning, given by the users' and developers' pre-understanding.

In Kim and March (1995), the same viewpoints are presented in a four-phase process model for requirements determination:

1. *Perception* – Users perceive the enterprise reality. The same enterprise reality may be perceived differently by different users (inconsistency). Any one of the users may perceive only a part of the reality (incompleteness).
2. *Discovery* – Analysts interact with users to elicit their perceptions.
3. *Modelling* – Based on the information identified in the discovery phase, analysts build a formal, conceptual model (representation) of the enterprise reality. This model serves as a communication vehicle between analysts and users.
4. *Validation* – Before concluding the model is correct, consistent, and complete, it must be validated. Validation has two aspects: comprehension and discrepancy checking. Users must comprehend and understand the meaning of the model. Then they must identify discrepancies between the model and their knowledge of reality.

Thus, in order to arrive at an inter-subjectively shared and agreed upon representation of the application domain, the user's concepts must be concretised and reconciled into a common vocabulary. This suggests that classification can be seen as part of the discovery phase, and as a prerequisite to the modelling phase. As part of the discovery phase, classification may collectively refer to both the process of classification as well as to the end result of the classification process. While the process is concerned with concept definitions and vocabulary construction, the end result is a common vocabulary to be used as input to the modelling phase. For definitions and in and further details about classification and related concepts, the reader is referred to chapter 2 and 4.

Similarly, conceptual modelling may be viewed as a process whereby the users' and developer's knowledge of the application domain is given a uniform and explicit representation, in the form of a conceptual model. This model, in turn, may be understood as a symbolic representation of the key concepts and relationships that make up the domain. For definitions and in dept analyses of conceptual modelling and related concepts, the reader is referred to chapter 4.

The fact that concepts are symbolically represented does not necessarily mean they are intensionally defined. On the contrary, as commented by Bergamashi and Sartory (1992), the idea of intensional definitions is almost unheard of in the conceptual model tradition. Rather, emphasis has normally been placed on the descriptive properties of objects. The definition of a class in conceptual modelling is generally understood as the definition of its descriptive properties. Since the definitional properties necessarily must be the same for all objects in a class, it might be that they are considered redundant, and hence excluded from the class definition.

As a consequence, the membership criteria, which are supposed to settle whether an object belongs to a class, will at best remain as a commentary in a data dictionary, and hardly ever be noticed during the design and implementation of the application. However, the definitional properties play a critical role in any class-based application for several reasons:

First, at the conceptual level, membership conditions should be represented by concept definitions. By using intensional definitions, the resulting system of concept definitions, or vocabulary for short, may be evaluated for its completeness and logical consistency. The vocabulary is complete when it includes all the concepts mentioned in the information requirements and logically consistent when all concepts are properly defined by intensional definitions. The hierarchical structures that result from intensional definitions may easily be checked for its logical consistency.

Second, the logical and hierarchical structures that result from the definition process may guide the naming, selection and justification of entity types, structural relationships and roles in the conceptual and logical models. Since the conceptual level is mainly concerned with the intensional aspects of concepts, design decisions at this level will be motivated by intents to make the model as simple as possible, yet rich enough to convey the meaning of the concepts. At the logical level, which is mainly concerned with extensional aspects, conceptual structures may be inflated or conflated due to inheritance considerations. At both levels, the vocabulary will provide a framework for discussing the design decisions that are made.

Third, membership conditions are the only means to control that objects that enter a class really belong there. If users are unaware the membership conditions for a class, incorrect instances may be recorded. Hence, for class-based applications, membership conditions should be formalized and controlled by the application. At the conceptual level, the membership condition can be expressed in natural language, as it would appear in the vocabulary. At the logical level, membership conditions may be operationalized and complemented by an algorithm for the actual checking that must be done. At the physical level, the procedure may be implemented by means of triggers, procedures or methods, depending on the chosen DBMS.

Fourth, problems related to homonymous and synonymous class terms are easily confused with differences in attributes. This is especially evident when attempts are made to integrate two or more separate applications. Since we usually consider types, we tend to assume that two classes are homonymous if they have the same class name, but differ in their attributes or synonyms if they differ in class names but have similar attributes.

However, by emphasizing the distinction between classes on one hand, and types on the other, it becomes evident that the objects that constitute a class may be variously described for different purposes, or from different perspectives, while still being of the same kind. Hence, with respect to schema integration problems, membership conditions should be among the first things to inspect

1.2 Related work

During the last decade, a number of theoretical papers have been published, in which key concepts such as: class, concept, membership condition, intension and extension, and classification have been of prime concern. Wand et al., (1995), propose a foundation for conceptual modelling based on ontology from Philosophy, classification theories from the Cognitive Sciences, and Speech Act theory from Linguistics. The fact that these ideas are further developed in successive papers by Parsons (1996), Parsons and Wand (1997a), Parsons and Wand (1997b), and Parsons and Wand (2000), is a clear indication of a current interest in discussing and advancing our understanding of the more fundamental aspects of conceptual modelling. There is also a well developed chapter in Martin and Odell (1992), in which all key concepts mentioned above are elegantly exemplified and discussed.

Hakim and Garrett (1997), suggest combining object-oriented modelling concepts with description logics, in order to overcome a number of limitations that follow directly from the inability of current object-oriented languages to define concepts by their necessary and sufficient conditions. Description logic is a kind of KR- language, which is divided into two separate languages: a terminological language to define concepts and relationships between concepts, and an assertional language, to create and manipulate individuals. The distinction between a terminological and an assertional language parallels our intuition that conceptual modelling should be similarly divided into definitional and descriptive parts.

Terminology is also a central issue in the current research on ontologies for knowledge-based systems. An ontology is considered a fundamental tool to support interoperability between knowledge systems, i.e., when knowledge sources are fused into a combined resource, like for instance a data warehouse, or when knowledge is to be shared among several knowledge-bases. Gamper, Nejdl, and Wolpers (1999) explore the commonalities and differences between ontologies and terminologies. Wand, Storey and Weber (1999), use ontology to analyse the meaning of common conceptual modelling constructs, and Guarino and Welty

(2000) present a formal ontology of properties, in which important distinctions between membership conditions, identity conditions, object identifiers and primary keys are discussed. Finally, my suggestion to explore the relevancy of various classification theories to conceptual modelling coincides with suggestions from Booch (1991), Wand, Monarchi, Parsons and Woo (1995) and Parsons (1996), who all introduce theories of classification from the cognitive sciences.

1.3 Assumptions and motivations

After more than 15 years of teaching data modelling and database application design to both undergraduate and graduate students, I have become more and more aware of, and frustrated by the fact that the discipline seems to lack a shared understanding of its basic concepts, such as concept, object, property, class, type, relationship, role, classification, generalization and inheritance. In order for the students to understand the meaning of the concepts and the subtle nuances that sets them apart, it has become customary for me to rework the textbook definitions, no matter which textbooks have been used. This is not an ideal situation, because it confuses the students and makes them question the overall quality of the textbooks, (or the teacher), all from the start.

Over the years I have found it useful to start with a definition of classification as a process whereby mental concepts are concretised and expressed by concept definitions. This definition requires a number of other concepts to be defined and distinctions to be made, for instance between defining and descriptive properties, classes and types, classification and identification. See chapter 2 and 4 for further details.

In my view, classification is a key concept from which it is possible to develop a coherent set of definitions for the basic concepts that pertain to conceptual data modelling. In addition, I am quite confident that both designers and reviewers of conceptual data models may benefit from the additional semantics that result from classification, i.e., the vocabulary of terms that a conceptual data model is based on. However, in spite of its assumed centrality and importance, I have a very firm impression that classification has received a rather marginal treatment in most textbooks. To find out whether this impression is correct, and to learn more about how, and to what extent classification may influence the design and interpretations of conceptual data models, a list of four major research objectives are given below.

1. To develop a coherent set of concept definitions, where the concepts pertaining to classification are clearly distinguished from, yet closely related to the concepts pertaining to conceptual data modelling.

2. To show that classification has not received sufficient attention with respect to conceptual data modelling.

3. To develop a method of classification, providing guidelines on how to perform and how to check the results of classification.

4. To study the effects of classification on the design of conceptual data models and on the interpretation of conceptual models.

1.4 Research questions and methods

The research questions are arranged according to the broad research objectives in the previous section.

1. To develop a coherent set of concept definitions, where the concepts pertaining to classification are clearly distinguished from, yet closely related to the concepts pertaining to conceptual data modeling.

   With respect to conceptual data modeling, what is the meaning of the following concepts:

   a. *Classification* versus *conceptual data modelling*?
   b. *A classification* versus *a conceptual data model*?
   c. *Classification* versus *identification*?
   d. *Concept* versus *type*?
   e. *Object* versus *entity*?
   f. *Property* versus *attribute*?

The research approach is based on concept analysis of existing theories of classification and conceptual modeling, most notably from the Cognitive sciences, Terminology, Archaeology, Conceptual Data Modeling, and from Knowledge Representation. For further details of the concept analysis approach, see chapter 2.

2. To show that classification has not received sufficient attention with respect to conceptual data modelling.

   a. How do current textbooks on data-oriented and object-oriented methodologies define classification, and related notions?

b. Which guidelines do current textbooks on data-oriented and object-oriented methodologies provide for classification and documentation of concepts?

The two research questions will be addressed by conducting a critical review of selected textbooks on data-oriented and object-oriented methodologies. The sampling procedure is further described in chapter 3.

3. To develop a method of classification, providing guidelines on how to perform, and how to validate the results from classification.

   a. How are concepts identified?
   b. How are concepts defined?
   c. How are concept definitions operationalized?
   d. How may the results from classification be validated with respect to its logical consistency?
   e. How may the results from classification be validated with respect to its completeness?
   f. How may the results from classification be used to validate a conceptual model?
   g. How may a conceptual model be used to validate the results from classification?

The methodological guidelines are derived from research question 1 and 2.

4. To study the effects of classification on the interpretation and the design of conceptual data models.

   a. How does classification affect the interpretation of conceptual data models?
   b. How does classification affect the design of conceptual data models?

Based on hypotheses derived from the concept analysis, an experimental research design is used to empirically test the effect of classification on interpretation tasks.

2.0 Concept analysis of classification

Concept analysis is a research method used to quantify and analyze the presence, meanings and relationships of concepts expressed in language. There are several approaches to concept analysis, stretching from set-theoretic methods operating on data sets, in order to discover dependencies within the data, to methods that use literature as data, in order to develop concepts within a particular discipline, cultural group, or the context provided by a particular theory. An example of the first kind is Formal Concept Analysis (Mineau, Stumme, and Wille, 1999), which has been applied in conceptual clustering, statistical classification, information retrieval, knowledge discovery, and ontology engineering. An example of the second kind is Evolutionary Concept Analysis (Rodgers and Knafl, 2000), which has been applied with a number of literature-based analyses of diverse concepts within the nursing discipline.

In this study, Evolutionary Concept Analysis has been selected for the following reasons: it focuses on concept development to solve conceptual problems; it uses literature as data; emphasis is placed on inductive inquiry and rigorous analysis; it supports inter-disciplinary and cross-disciplinary analyses; it leads to the generation of implications and hypotheses about the pragmatic utility of the results, and provides a basis for further inquiry by whichever methods the researcher finds necessary. In addition, it is based on current philosophical thought rejecting essentialist ideas of isolated, finite, concept definitions, in favour of conceptual change. The emphasis on *evolution* and *development* in the name of the method is deliberately used to reflect the idea of conceptual change.

> “The emphasis on conceptual change points to the idea that concept development must be an ongoing process, with no realistic end point, except that work on a concept decrease as the concept looses significance. As phenomena, needs, and goals change, concepts must be continually refined and variations introduced to achieve a clearer and more useful repertoire. Attempts to delineate precise or definitive boundaries, to distinguish a concept from its context, or to view it apart from a network of related concepts, as often done with concept analysis, are not consistent with this view” (Rodgers and Knafl, 2000, p. 82).

The method is considered to be well suited for the current study, which uses scientific literature as data. In accordance with the method’s inductive approach, the literature sample is collected from several disciplines and sub-disciplines, requiring both cross disciplinary and interdisciplinary analyses. In addition, the method’s heuristic function that leads to the identification of directions for further inquiry makes a smooth transition from the current study to hypotheses testing in subsequent studies as reported in chapter 3 through 6.

The philosophical basis on which the method is grounded is also very much in line with my own view on concepts and conceptual modelling. I believe that concepts are private, subjective and dynamic constructs that must continuously adapt to changes in theoretical knowledge, goals and requirements from within the context in which they are used. However, within disciplinary domains, or formal contexts such as particular application domains, the need to ensure that conceptual frameworks are consistent across individuals requires concepts to be concretized and formalized to various degrees. In such situations, individuals may adjust their conceptions and come to agreement on a unified interpretation of the framework.
This study is an attempt to develop a conceptual framework that clarifies the concept of classification in the context of conceptual modelling. In accordance with the philosophical view just outlined, the concepts and relationships so developed may have utility within the context of conceptual modelling, and not necessarily outside that context. In addition, since concepts are considered to be dynamic construct, the framework does not represent finite definitions, but rather a contribution to an ongoing process of conceptual change in the discipline.

2.1 Purpose.

The purpose of this study is to develop a conceptual framework that clarifies the concept of classification within the context of conceptual modeling. The term conceptual framework is used here to mean a meaningful and elaborate system of concepts that can be used to describe and reason about concepts or phenomena, to reveal new insights, to provide directions for research, and to point at solutions to problems. In order to evaluate its pragmatic utility the framework should be useful in the following respects:

1. Clearly reflect the meaning of classification as it pertains to conceptual modeling.
2. Provide guidelines on how to use classification in conceptual modeling.
3. Provide guidelines on how to validate the results of classification.
4. Contribute to the development of a coherent vocabulary for classification and conceptual modeling.

In addition, hypotheses about the pragmatic utility of the framework will be generated and made subject to subsequent inquiries.

2.2 Method.

The evolutionary method of concept analysis contains 6 steps:

1. Identify the concept of interest and associated expressions (including surrogate terms).
2. Identify and select an appropriate realm (setting and sample) for data collection.
3. Collect data relevant to identify:
    a. The attributes of the concept; and
    b. The contextual basis of the concept, including interdisciplinary, sociocultural, and temporal (antecedent and consequential occurrences) variations.
4. Analyze data regarding the above characteristics of the concept.
5. Identify an exemplar of the concept, if appropriate.
6. Identify implications, hypotheses, and implications for further development of the concept.

The 6 steps represent tasks to be accomplished rather than a specific, fixed sequence of steps in a process. Steps may be iterated or carried out simultaneously as the investigation proceeds.

2.2.1 Identification of the concept of interest and associated expressions.

In accordance with the inductive approach to identification, no preconceived ideas of classification were used to delimit the initial search space. Hence, a free text search for conference proceedings and scientific articles was carried out based on a set of broad terms, along with associated "surrogate terms" as shown in table 2.1, 2.2 and 2.3 below.

Construct terms	Process terms
Concept	Classification
Property	Modeling
Class	Testing
Classification	Integration
Model	

Table 2.1: Main search terms

Based on these five construct terms and four process terms, two surrogate tables were gradually developed, containing terms that were used interchangeably to denote the same or related concepts. The two tables were continually expanded by new terms as the identification process proceeded.

Concept	Property	Class	Classification	Model
Abstraction	Attribute	Group	Concept System	Data Model
Idea	Feature	Aggregate	Categorization	Datamodel
Class	Dimension	Category	Generalization	Conceptual Model
Category	Value	Class Definition	Specialization	Conceptual Data Model
Term	Data	Data Definition	IS_A	Information Model
Name	Data Source	Domain	Taxonomy	Semantic Data Model
Data Name	Description	Entity Type	Typology	Enterprise Model
Class Name		Extension	Abstraction	Corporate Data Model
Terminology		Extensional	Hierarchy	Logical Data Model
Vocabulary		Object Type		Physical Data Model
Definition		Entity Type		Relational Model
Intension		Subclass		Conceptual Schema
Intensional		Superclass		Logical Schema
Nomenclature		Taxon		Physical Schema
Object		Type		Distributed Databases
Entity		Set		ER Model
		Representation		Entity Relationship Model
				Semantic Data Model
				SQL, SQL3, OQL, DDL
				Data Definition Language
				Meta Data Model
				Ontology
				UML, OSADM, OOAD, CG, ...
				Data Catalogue
				Meta Data
				Metadata
				Description Logics

Table 2.2: Surrogate table for construct terms.

Classification	Modeling	Testing	Integration
Grouping	Modelling	Valid	Database Integration
Categorization	Analysis	Validity	Application Integration
Categorize	Analysis and Design	Validation	Schema Integration
Identification	Domain Modelling	Evaluation	Database Evolution
Identify	Domain Analysis	Integrity	Schema Evolution
Generalization	Conceptual Modelling	Coherency	Data Sharing
Generalize	Conceptual Data Modelling	Coherent	Homonym
Specialization	Conceptual Analysis	Consistency	Synonym
Specialize	Semantic Data Modelling	Consistent	View Integration
Abstraction	Semantic Analysis	Data Quality	Interoperability
Abstract	Information Modelling	Data Cleaning	Database Mapping
Aggregation	Enterprise Modelling		Schema Mapping
Aggregate	ER Modelling		View Mapping
Definition	Design		Transformation
Define	Logical Modelling		Translation
Inheritance	Logical Design		Cooperation
	Knowledge Representation		
	Software Engineering		

Table 2.3: Surrogate table for process terms.

The identification process was based on published studies in proceedings, journals, and textbooks that were available for searching and loan ordering via BIBSYS. BIBSYS is a shared, online library system for all Norwegian University Libraries, the National Library and a number of college and research libraries. In addition to its holding database, which contains bibliographic data about 8.0 mill documents, BIBSYS also has a citation database based on data from the Institute for Scientific Information (ISI). This database provides access to current and retrospective bibliographic information, author abstracts, and cited references about 14.2 mill articles, published in 5,800 of the world's leading scientific and technical journals, 1,700 of the world's leading social sciences journals, and over 1,400 of the world's leading arts and humanities journals.

During the identification process, new ideas, current research issues, as well as more established knowledge were considered as relevant. Accordingly, the most recent proceedings were systematically reviewed in order to capture the latest research ideas. In addition, advanced article searches were carried out to capture current research, as well as established knowledge. Finally textbooks were reviewed, with a special focus on well-established knowledge.

As documents were selected, the lists of surrogate terms were continuously expanded by inclusion of keywords supplied by the authors, as well as keywords supplied by ISI. Hence, new keywords were used in subsequent search processes, along with combinations of terms that were found to be too broad during the initial search.

As the identification process proceeded, it became clear that documents that were concerned with cognitive, representational, or practical/theoretical aspects of classification were most relevant to understand classification in the context of conceptual modelling. Cognitive aspects of classification are concerned with how people conceptualize the world, how mental concepts are learned and used. Representational aspects are concerned with symbolic representations of knowledge, and ways to concretize mental representations.

Practical/theoretical aspects covered general ideas of classification, principles, classification structures, historical, philosophical and metaphysical reflections on the topic.

In the end, a search process was performed based on references to persons such as keynote speakers at conferences, and authors of invited papers. Accordingly, their names were used in a subsequent search by author, in order to list and review their publications.

The search for proceedings was performed using the search term “Proceeding?” in the title-field, combined with search terms from table 2 and 3 in the free text search field. The *free text search option* searches the database(s), in this case the BIBSYS holding database and the ISI citation databases, for matching terms in the title, the abstract, and the keywords that are supplied by the author, or by ISI. If a match is found the document is listed by title, author, year of publication, and type.
The list of proceedings was then manually reviewed with respect to the titles and date of publication. The most recent and relevant proceedings were selected for further reviews. Selected conference proceedings are listed below.

International Conference on Conceptual Modeling, ER ’99, 2000
International Conference on Knowledge Engineering and Knowledge Management, EKAW 2000
International Conference on Conceptual Structures ICCS 1993, 1999, 2000
International Congress on Terminology and Knowledge Engineering, TKE ‘99
IFIP International Conference on Information System Concepts, 2000
International Conference on Information and Knowledge Management, CIKM 2000
ACM SIGSOFT Sixth Int Symposium on the Foundation of Software Engineering, FSE-6
International Conference on Object-Oriented and Entity-Relationship Modeling, OOER ‘95
Conceptual Modeling – Current issues and Future Directions (1999)
European Workshop on Knowledge Acquisition, Modeling and Management, EKAW 1999
International Conference on Advanced Information Systems Engineering, CAiSE’96

A similar process was used to search for articles in the ISI citation databases, but then, only the free text search field was used. Some terms were truncated in order to capture different spellings, (e.g., “model?” as a substitute for “model”, “models”, “modeling” and “modelling”). Other terms were combined and split, (e.g., “metadata” and “meta data”).
If the retrieved list of documents exceeded 500, the list was discarded, and the search term(s) marked for subsequent use in combination with other terms (e.g., “data” and “model” combined into the new term “data model?”). Any list that contained less than 500 entries was reviewed with respect to the titles. Entries, for which the title seemed relevant, were further reviewed with respect to its abstract.

2.2.2 Selection of setting and sample.

During the identification process, a total of 288 documents were selected for inclusion in the study. Each document was numbered sequentially and a sample of 115 documents (N=288, n=115) was selected by means of computer-generated random numbers. The sample size of 115 documents equals 40% of the total collection. According to Rodgers (2000), 20%, or at least 30 papers are considered as a minimum to facilitate a credible analysis. However, because of the interdisciplinary nature of classification, the percentage was doubled in order to obtain an acceptable coverage of cognitive, representational and practical/theoretical aspects of classification. In addition to the randomly generated sample, a selection of papers considered to be classic, specially invited, or surveys were added, increasing the total sample size to n=127. These include papers by Abrial (1974), Chen (1976), Codd (1979), Bubenko (1980), Hammer and McLeod (1981), Murphy and Medin (1985), Medin (1989), Hempel (1994), Gruber (1995), and Mylopoulos (1998).

Sorted by content the articles gave the following gross distributions:

Topic	No of articles
Cognitive aspects of classification	16
Representational aspects, including representational languages and modelling approaches	66
Practical and theoretical aspects, including principles and techniques, taxonomies, typologies	31
Schema integration	14

Table 2.4: Gross distribution of articles sorted by content.

The numbers may give the impression that cognitive aspects are underrepresented, but at least 17 of the documents in the representational aspects category could just as well be categorized as belonging to the cognitive aspects category. Similarly, at least 11 documents from the representational aspects category could easily have been categorized with the practical and theoretical aspects category.

As for the schema integration category, 9 more articles from the remaining collection were added, and a new electronic search was made in order to compile a minimum sub-sample of 30 papers, giving a total sample of 143 papers. Table 2.5 shows how the papers are distributed on scientific journals.

Journal	**No of articles**
American Antiquity	3
ACM Transactions on Database Systems	6
Cognitive Psychology	2
Communications of the ACM	4
Data and Knowledge Engineering	7
Information and Software Technology	4
Information Modelling and Knowledge Bases	5
Information Systems	7
Int. Conf. on Conceptual Modeling	10
Int. Conf. on Knowledge Organization and Quality Management (ISKO)	2
Int. Journal – Human-Computer Studies	5
Knowledge Organization	2
Minds and Machines	2
Miscellaneous	84

Table 2.5: Distribution of articles sorted by Journal.

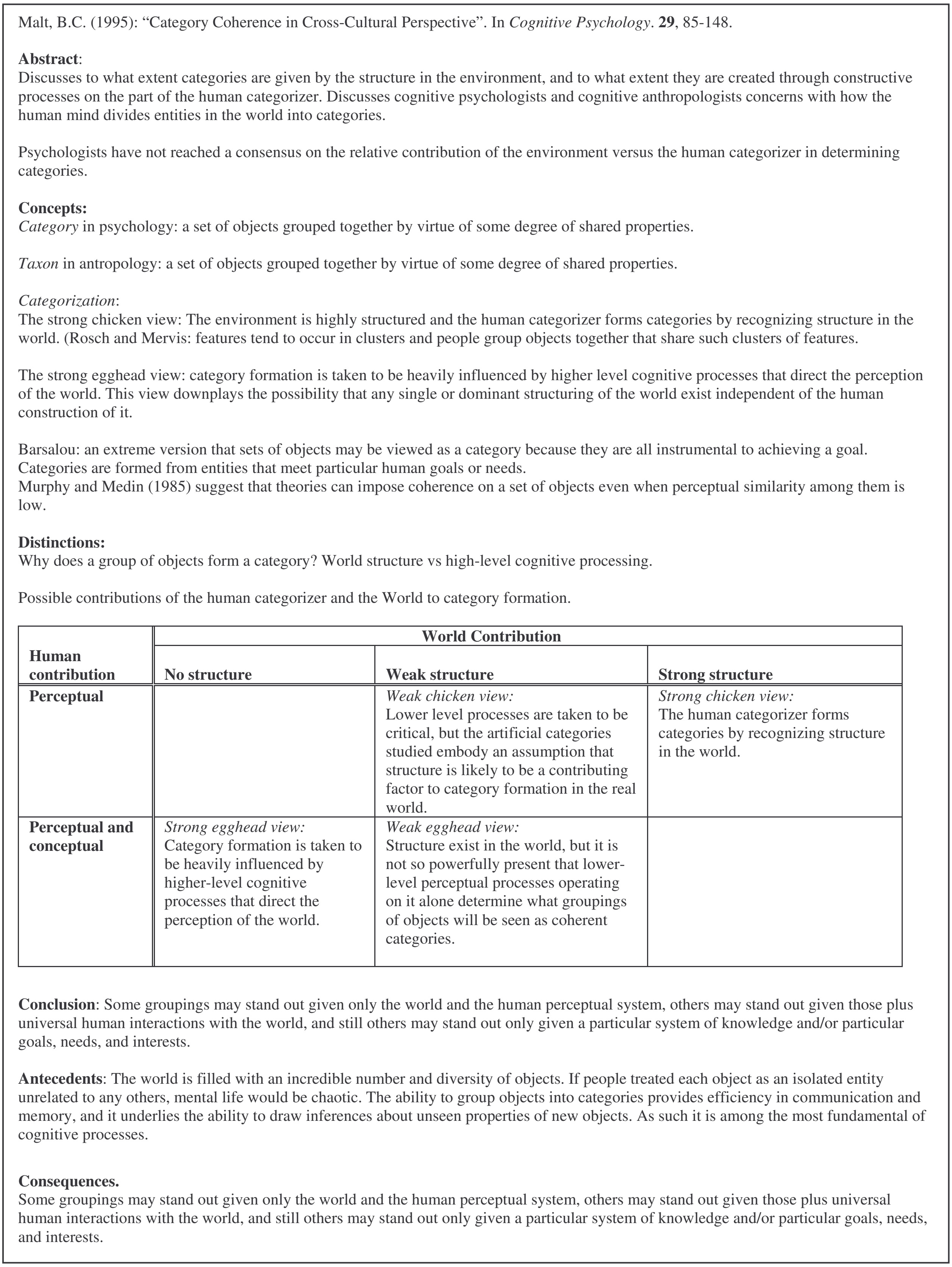
2.2.3 Data collection

As mentioned in section 2.2.1, the literature was obtained through a shared, online library system. As the papers were received, an initial, minimal analysis was conducted in order to identify new search terms, to provide directions for further investigation, and to suggest an efficient organization of data to facilitate the analysis. Table 2.2 through 2.5 are intermediate results from this initial analysis.

Prior to the actual data collection, the papers were sorted into four piles according to the topics in table 2.4. The papers were then reviewed several times, and relevant data were underlined or commented directly in the papers. After the first review, it was decided to postpone the analysis of papers concerning schema integration, leaving the papers on cognitive, representational, and practical/theoretical aspects of classification for the analysis. The actual data were collected from repeated reviews and relevant data were recorded about:

a) the attributes of the concept, i.e. its defining characteristics.
b) its contextual features, such as antecedents, consequences and disciplinary contexts.
c) surrogate terms, i.e., other terms or means of expressing the concept.
d) related concepts, that may help to situate the concept in a broader knowledge structure.
e) applications of the concept, i.e., how it is used.
f) developmental perspectives that portray changes of the concept over time.

The data from each paper were recorded on separate sheets. In addition, thoughts and ideas, as well as cross references were added to the sheets with separate entries. An example sheet is shown in figure 2.1 on page 17.



Malt, B.C. (1995): "Category Coherence in Cross-Cultural Perspective". In *Cognitive Psychology*. **29**, 85-148.

Abstract:
Discusses to what extent categories are given by the structure in the environment, and to what extent they are created through constructive processes on the part of the human categorizer. Discusses cognitive psychologists and cognitive anthropologists concerns with how the human mind divides entities in the world into categories.

Psychologists have not reached a consensus on the relative contribution of the environment versus the human categorizer in determining categories.

Concepts:
Category in psychology: a set of objects grouped together by virtue of some degree of shared properties.

Taxon in antropology: a set of objects grouped together by virtue of some degree of shared properties.

Categorization:
The strong chicken view: The environment is highly structured and the human categorizer forms categories by recognizing structure in the world. (Rosch and Mervis: features tend to occur in clusters and people group objects together that share such clusters of features.

The strong egghead view: category formation is taken to be heavily influenced by higher level cognitive processes that direct the perception of the world. This view downplays the possibility that any single or dominant structuring of the world exist independent of the human construction of it.

Barsalou: an extreme version that sets of objects may be viewed as a category because they are all instrumental to achieving a goal. Categories are formed from entities that meet particular human goals or needs.
Murphy and Medin (1985) suggest that theories can impose coherence on a set of objects even when perceptual similarity among them is low.

Distinctions:
Why does a group of objects form a category? World structure vs high-level cognitive processing.

Possible contributions of the human categorizer and the World to category formation.

	World Contribution		
Human contribution	**No structure**	**Weak structure**	**Strong structure**
Perceptual		*Weak chicken view:* Lower level processes are taken to be critical, but the artificial categories studied embody an assumption that structure is likely to be a contributing factor to category formation in the real world.	*Strong chicken view:* The human categorizer forms categories by recognizing structure in the world.
Perceptual and conceptual	*Strong egghead view:* Category formation is taken to be heavily influenced by higher-level cognitive processes that direct the perception of the world.	*Weak egghead view:* Structure exist in the world, but it is not so powerfully present that lower-level perceptual processes operating on it alone determine what groupings of objects will be seen as coherent categories.	

Conclusion: Some groupings may stand out given only the world and the human perceptual system, others may stand out given those plus universal human interactions with the world, and still others may stand out only given a particular system of knowledge and/or particular goals, needs, and interests.

Antecedents: The world is filled with an incredible number and diversity of objects. If people treated each object as an isolated entity unrelated to any others, mental life would be chaotic. The ability to group objects into categories provides efficiency in communication and memory, and it underlies the ability to draw inferences about unseen properties of new objects. As such it is among the most fundamental of cognitive processes.

Consequences.
Some groupings may stand out given only the world and the human perceptual system, others may stand out given those plus universal human interactions with the world, and still others may stand out only given a particular system of knowledge and/or particular goals, needs, and interests.

Figure 2.1: Example data collection sheet with cross references to other writers and with a table which presents various views on the antecedents to classification.

2.2.4 Analysis.

Based on recommendations in Rodgers and Knafl (2000), the final analysis was delayed until the data collection was almost complete. The reason for this was to avoid any premature closure or tendencies to seek confirmation on preconceived notions.

During data collection the data sheets were assembled in three separate documents, one for the cognitive aspects, one for representational aspects, and one for practical and theoretical aspects. Consequently, the analysis was intended to follow the same division. However, during the analysis, a different structure emerged from the data.

First, it became clear that several disciplines had experienced a methodological debate, where ontological, epistemological and methodological aspects of classification were of central concern.

Second, among the practical and theoretical papers, the contributions from disciplines such as Terminology and Archaeology were so convincing and well articulated that they deserved a position on par with the Cognitive and Representational disciplines. As a result, metaphysical aspects of classification were first analyzed for all disciplines. Then, each discipline was analyzed one at a time. In the end, the various perspectives were compared, and attempts were made to generalize from the findings. The final structure resulting from the analysis is reflected in section 2.3.

For each discipline, the various entries on the data sheets were compared and contrasted, and the results were organized and reorganized several times. Very often the papers would have to be consulted again as new insights suddenly made inaccessible parts of the texts comprehensible. To identify similarities and differences, definitions and terms were organized in tables, and separate notes were continuously made, adjusted, and refined. As a system of concepts emerged from the analysis, tentative definitions were scrutinized and evaluated for their consistency and coherency. Their respective utility was also considered with respect to the requirements set out in section 2.1. In addition, the concepts were tested with respect to their place and use in a methodological framework for conceptual modelling. To see how the concepts fit in with the method, see chapter 4.

Several measures have been taken to strengthen the credibility and rigor of the current analysis. For a full discussion, see section 6.3.1 on page 145.

2.3 Findings.

In general, classification is a term with at least three different, but related senses, (Sokal, 1974). First, classification is used to mean a process of defining classes. Second, classification is used to denote the system of classes that result from the classification process. Third, classification is used to refer to the judgment that must be exercised, in order to assign a particular thing to its proper class.

Although there is a general agreement on these senses, disagreement appears once one starts to ask about the nature of the classes and the classification systems so created. Do they reflect natural divisions that exist in the world, or are they simply arbitrary structures to suit our needs and purposes? These questions have been dealt with since the time of Plato and are known as the Problem of Universals, which is concerned with whether there are universals, and what it is that the general terms in our language refer to. Most of the solutions that have been suggested fall under one of three broad views, called realism, nominalism, and conceptualism, (Kangassalo, 1992; Audi, 1995; Artz, 1997; Mylopoulos, 1998; and Lane, 2002)

2.3.1 Metaphysical perspectives on classification

Realist perspectives.

According to realists, singular terms refer to particulars, while general terms refer to general objects, called universals. Particulars are the individual objects that can be encountered in the world. They are characterized by being spatial, temporal, transient, changeable and singular. They have properties and they enter into relations independently of the concepts with which we understand them, or of the language with which we describe them.

Universals, on the other hand, are considered as abstract objects such as properties, relations, numbers, and laws of logic and nature. As opposed to particulars, universals are characterized by being non-spatial, timeless, general, unchangeable, and necessary. Both particulars and universals exist independently of our experience or our knowledge of them.

Because of the generality, stability and necessity of universals, it is commonly held that universals, particularly properties, serve a classificatory function by representing real and invariant structures in the world. Properties, therefore, are understood as the principles of classification, which a person either knows, or of which he is ignorant, or about which he has false beliefs.

Accordingly, realists hold that that our classification systems are determined by a reality, which is independent of us, and that the classification process is a matter of discovery.

Nominalist perspectives.

According to nominalists, singular terms refer to particulars, while general terms refer to collections of objects. Universals do not exist. Individual objects are the primary existents, and properties are considered as distinct and inseparable aspects of those individual objects. A property is not something that may be shared between objects, as realists hold, but rather something distinct to the object which possesses it.

In the most extreme form of nominalism, the only thing held in common by all the instances of a general term, is the general term itself. Since no two objects can have any properties in common, the application of a general name to one object, rather than another, becomes arbitrary and subjective. There are no kinds to which a thing belongs, no common properties to serve any classificatory functions, and hence no basis in reality for our classification systems. Consequently, nominalists hold that our classification systems are determined by a social consensus and/or social conventions on the use of general terms, and that the classification process is a matter of linguistic analysis.

Conceptualist perspectives.

Conceptualism, sometimes also called moderate realism, can be regarded as a resolution between nominalism and realism, where abstract, mental concepts are introduced to mediate between general terms and objects. According to conceptualists, universals exist, but only as abstract concepts in the mind. General terms refer to concepts in the mind, and the concepts refer to objects in the real world. This view accords with Aristotle's view of universals, in which universals exist, but only, insofar as they are instantiated in specific things. According to Aristotle, we have knowledge of two different kinds of objects. The senses give us awareness of particular and concrete things around us, while the intellect has the capacity to form and reason about abstract concepts. These concepts are formed through the process of abstraction, which is an intellectual process of recognizing the commonalities among a number of objects.

Accordingly, conceptualists hold that the supposed classificatory function of universals is served by our mental concepts, which are constructed by contributions from our intellect, and from the objective structures of the environment.

To conceptualists therefore, classification systems are determined by our mental concepts, and the classification process becomes a matter of abstraction, which yields results that are probable, but not necessarily true.

2.3.2 A framework for analysis of ‘classification’

When considering the various positions on the problem of universals, three related concepts stand out as especially important to include in a concept analysis of classification. Those are *concept*, *term*, and *class*. The concepts and their relationships may be visualized in a simple analysis framework based on Ogden and Richards (1972) classical meaning triangle.

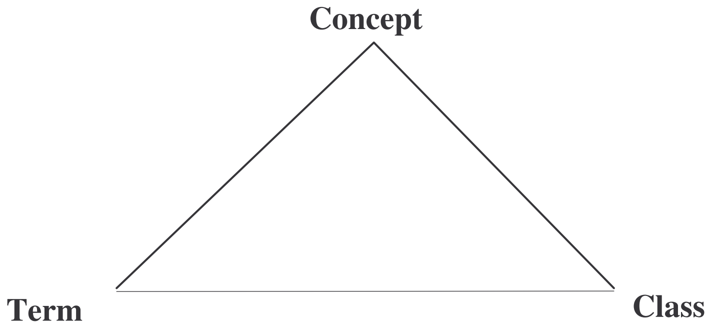


Figure 2.2: Analysis framework for classification and related notions.

In what follows, cognitive, practical/theoretical, and representational perspectives on classification will be presented and analyzed with reference to the analysis framework above.

2.3.2.1 Cognitive perspectives on classification.

In order to understand how people conceptualize the world, cognitive psychologists have focused extensively on classification, concepts, and classes. The preferred terms used by most psychologists, however, are ‘categorization’, ‘concept’, and ‘category’. Although there is a general agreement that categorization is a fundamental cognitive process, and that concepts are mental constructs, it is hard to find a consensual view on the *sources* of conceptual order, or on what concepts and categories really are.

According to Malt (1995), neither psychologists nor anthropologists have reached a consensus on the relative contribution of the environment versus the human categorizer in determining categories.

While some hold that the environment is highly structured and that the categorizer form categories by directly recognizing structure in the world, others hold that category formation is heavily influenced by cognitive processes that direct the perception of the world.

When it comes to controversies about concepts and categories, Hampton and Dubois (1993) have pointed out that cognitive science is a relatively young scientific discipline where rival theories of categorization have led to a certain degree of incommensurability between terms. Similarly Van Mechelen, Boeck, Theuns and Degreef (1993) claim that controversies about the definition of concept and category leave the observer with the impression that much energy has been invested in various falsification attempts. Consequently, it is more clear what categories and concepts are not than what they are. This will become evident as the theories are briefly surveyed below.

Since the establishment of cognitive science in the 1950s, five major psychological theories of classification have been advanced: the classical theory, the prototype theory, the exemplar theory, the frame view and the theory view.

The classical theory.

According to Lakoff (1987) the classical theory is derived from a philosophical position arrived at on the basis of *a priory* speculation. The theory, which dates back to Plato and Aristotle, suggests that the world exhibits a universal taxonomic order, in which particulars belong to natural classes, which are related by strict genus-species relations, forming a natural, hierarchical structure.

Over time, a logical version of the classical theory has developed, where a concept is understood as a *term*, which is associated with an *intension* and an *extension*. The *intension* consists of one or more defining properties that constitute a condition for becoming a member of the associated extension. The *extension* amounts to the class of objects within a universe of discourse that satisfy the membership condition contained in the term's intension. In other words, all classes are characterized by a membership condition that any object must meet in order to become a member of the class. As a psychological theory one can easily recognize its classical roots by the definition of concept and category:

> "A concept is a (1) relatively fixed, (2) generic, (3) mental representation, (4) consisting of a list (5) of defining features of a corresponding category, which is a (6) set of elements (called category members) distinguished in an all-or-none fashion from a complementary set of elements (called non-members)", (Mechelen, Boeck, Theuns, and Degreef (1993), p. 334).

However, as pointed out by the authors, each of the six numbered properties is contested by at least one of the other theories.

First, the classical theory came under attack because people were generally unable to account for the *defining properties* for a wide range of ordinary concepts, such as game, and number, thinking and work. Second, *borderline cases* were demonstrated, in which people were unable to decide whether a given instance belonged to a class or not. Third, *typicality effects* were demonstrated, in that people rated some members of a category as more typical members than others.

See (Barsalou, 1987; Medin, 1989; Hampton, 1993; and Hahn and Chater, 1997) for details about the insufficiency of the classical theory, and for simple surveys of competing theories.

The prototype theory.

The first attempts by cognitive science to address the apparent shortcomings of the classical theory, suggested the use of *polythetic* definitions. A polythetic definition consists of a list of properties, none of which needs to be necessary. To become a member of a polythetic class an object must have a sufficient number of properties in common with the other members of the class. The notion of polythetic definitions forms the basis for both the prototype theory, and the exemplar theory.

The prototype theory holds that a concept is a mental construct that represents the prototypical properties for its category members. Prototypical properties are the ones that are most frequently occurring in members of the category. So, whenever an object is encountered, it is classified as a member of that category, with which it has most properties in common.

The exemplar theory.

Another polythetic variant is the exemplar theory, which claims that concepts are represented by a mental category of remembered exemplars. Rather than being compared to a single prototype, new objects will be compared to categories of remembered exemplars and classified with the exemplars to which it has greatest similarity.

With polythetic definitions, the problems with defining properties vanished. Polythetic definitions also seemed to explain the extensional vagueness of borderline-cases, since a polythetic approach will divide the universe into clusters of similar instances, each cluster having a well-defined center, while the border between one cluster and the next may be relatively poor. Finally, typicality effects could also be explained, since some instances would have more properties in common than others.

However, the three theories were soon to be challenged for their one-sided focus on properties and for their reliance on similarity as a means to explain why we have the categories we have and not others, (Storms and De Boeck, 1997).

Regarding properties, research has shown that people have substantial amounts of complex knowledge for familiar concepts which property lists fail to capture, (Barsalou, 1992). When researchers ask people to list all the relevant properties of a concept, they produce a tremendous amount of properties, correlated properties, parts, compositions, superordinates, subordinates, origins, related objects, operations, actions, functions, beliefs, frequency and so forth.

Regarding similarity, it is argued that the three theories fail to provide any account of why some similarities matter while an indefinitely large number of others do not, (Murphy and Medin, 1985; Medin, 1989). Similarity is simply too flexible to explain conceptual coherence: any two objects can be arbitrarily similar or dissimilar by changing the criterion for what counts as a relevant property. Hence, similarity is only useful to the extent that principles determining what is to count as a relevant property are specified. Such principles are assumed to be given by the background knowledge or naïve theories that people have about the world.

The frame view and the theory view.

In response to the apparent insufficiency of property lists and unconstrained similarity matching, the *frame view* and the *theory view* were proposed. Both theories share a common concern for complex conceptual structures. The frame view is concerned with the formalisms needed to *represent* such complex structures, while the theory view is concerned with the *content* and *formation* of those structures. According to the theory view both concepts and categories are seen as mental constructs that serve as building blocks for human thought and behavior. Concepts may not have real world counterparts, e.g. unicorns, and people may impose rather than discover structure in the world, e.g. goal-derived categories. For further details, see (Murphy and Medin, 1985; Lakoff, 1987; Medin and Wattenmaker, 1987; Medin, 1989; and Barsalou, 1992).

An important point to notice about the frame view and the theory view is that they mark a shift in the orientation away from a narrative and/or directly perceptual account of categorization toward a more theoretical and inferential basis: In McCauley's (1987) words:

> “Contrary, then, to empiricist learning theory, it is not from the blooming, buzzing confusion that we induce our categories, but rather from our idealizations that we impose them” (McCauley, 1987, p. 293).

This does not mean that all cognitive psychologists over time have turned to conventionalism and constructivism. Rather, there exist a whole range of views, stretching from pure realist to pure constructivist views, as well as mixtures of both.
Table 2.6 gives a simplified summary of the main theories and their key concepts.

Theories	Classical theory	Prototype theory Exemplar theory	Frame view Theory view
Categories are given by	objectively existing structures in the world.	world structure plus low-level perceptual processes.	high level cognitive processes that direct the perception of the world.
Concept	Mental representations of monothetic categorization rules.	Mental representations of polythetic categorization rules.	Mental constructs based on theories, models, goals, needs, interests.
Category	Class of real world objects according to real divisions in the world.	Class of real world objects according to probable divisions in the world.	Mental constructions. Explanatory relations between theories and the world.
Conceptualization	Discovery based on direct awareness of universal, organizing principles.	Discovery based on abstraction from sense-impressions.	Construction based on generic and episodic knowledge.
Categorization/ Identification	Exact attribute matching	Attribute matching	Inference process

Table 1.6: Key concepts in psychological theories of classification.

The way the theories are organized in the table does not suggest an either, or situation, where one view necessarily excludes the others. Though there are radical realists, such as Sutcliffe (1993) who hold that concepts are nothing but universal, organizational principles of an objective, material reality, most cognitive psychologists seem to accept the idea that there are different kinds of concepts and categories, such as logical, fuzzy, natural, abstract, and artificial ones, (Medin, Lynch and Solomon, 2000), some of which are best accounted for by the classical theory, while others are better accounted for by one of the other theories.

Based on the selected literature, the general view held by most cognitive psychologists seems to be somewhat like this: the world is filled with an incredible number and diversity of objects. If people treated each object as an isolated entity, mental life would be chaotic.

Therefore, there is a strong, human urge to organize the world of experience by creating orders, (Jacob, 1994; and Malt, 1995). This process of creating orders is seen as a fundamental cognitive process, called categorization, by which concepts are used to group objects into categories.

But to some researchers, concepts do more than just categorize. According to literature reviews made by Solomon, Medin and Lynch (1999), a concept is taken to mean a mental construct that is supposed to serve a variety of cognitive functions, such as categorization, learning, reasoning, explanation, prediction, and communication. Their research indicates that the notion of concept has gradually changed from that of being a relatively simple categorization rule, to some kind of knowledge representation that embodies a theory about the world. As a result of this conceptual development, different writers have developed their own understanding and use of terms, to the extent that there is currently no single, uniform view on the fundamental concepts within the field.

To the extent that one can speak of a general agreement on issues of classification, it must be that 1) there is a human need to organize the world, 2) categorization is a fundamental, cognitive process to do so, 3) concepts are mental representations, and, 4) concepts are subjective and liable to change both between subjects and, over time, within the same subject, depending on purpose, context and prior experience, (Barsalou, 1987). This flexibility of concepts have made Jacob (1994) point to an important distinction between categorization and classification. Categorization and classification are viewed as two different mechanisms for establishing order.

Categorization is a cognitive process of constructing order out of individual, day to day experiences and sense impressions. Because mental concepts are constructed to reflect the individual's encounters with the environment, they must be flexible and capable of responding both to the immediacy of experience and to the discovery of new patterns of similarity depending upon the context.

Classification, on the other hand, is a social process of structuring a specific knowledge domain, in order to ensure consistency and stability of meaning. To facilitate communication, mental concepts must be concretized in order to be talked about, negotiated, and shared. According to Jacob's view, categorization appears to be the justification for classification. It is exactly the vagueness, instability, and subjectivity of mental concepts that cognitive theories of categorization attempt to explain, and classification attempts to overcome.

This is further detailed in the next section where classification is treated from a practical/theoretical perspective, with contributions from terminology and archaeology.
Using the analysis framework from figure 2.2, the fundamental concepts with which cognitive psychology studies and explain classification can be presented in this way:

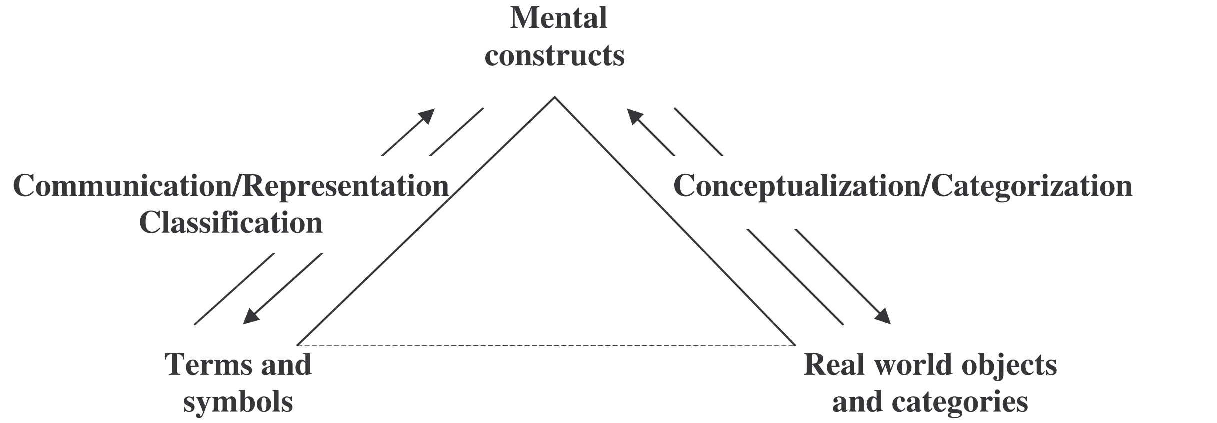

Mental constructs
Communication/Representation Classification
Conceptualization/Categorization
Terms and symbols
Real world objects and categories

Figure 2.3: Classification and related concepts from the perspective of cognitive psychology.

Cognitive psychologists are concerned with how people conceptualize, represent and use mental constructs to categorize, describe, understand, reason about, and communicate about the world. Instead of using the term 'mental concept' the term 'mental construct' has been deliberately chosen as a collective term for different kind(s) of mental representations. Instead of constantly expanding the meaning of 'concept' to include everything that matters to cognitive psychology, one could possibly benefit by restricting 'concept' to mean a mental construct that is used to support categorization, and introduce other terms to designate constructs that are needed to support other cognitive functions.
The nature of the relations between mental constructs and the world depends on which theory one adheres to, and the various interpretations are detailed in table 2.6.
The relations between mental constructs and language are many-to-many. In any natural language and in day to day conversation, there are no necessary connections between language and ideas. Connections are contingent and affected by the context, and the intentions of the subjects involved. For special languages however, this condition is different. In specific knowledge domains the need for consistency and stability of meanings requires concepts to be concretized by means of definitions.

This is what Jacob (1994) termed classification as opposed to categorization. By looking at the relations between concepts and language from a classification point of view, the relations are supposed to be restricted to one-to-one. This will be further detailed in the next section.

2.3.2.2 Practical/theoretical perspectives on classification.

Terminological perspectives.

Terminology is a term that designates both a discipline and its subject field. As a discipline, terminologists study concepts and their representations in terminologies, for the purpose of producing terminologies. Terminologies are the special languages used by experts and professionals in specialized domains, such as nursing, aviation, banking, geography, and so on. According to the International Standard No 704, Terminology work – Principles and methods, ISO 704:2000(E), terminology is multidisciplinary and draws support from a number of other disciplines, such as logic, epistemology, philosophy of science, linguistics, information science and cognitive science. The concepts most fundamental to terminology are *objects*, *concepts*, *designations*, and *definitions*.

Objects are considered material, such as computers, trees, or stars, or immaterial, such as numbers, the unicorn, or the first female Pope. One central assumption is that once objects are perceived as meaningful units of thought, their common properties are abstracted as characteristics, which are combined to create the concept. A characteristic is a generalization of a set of object properties. Color, for example, may be abstracted as a relevant characteristic from a set of *yellow*, *green*, and *red* objects.

Concepts are considered as mental representations of objects within a specialized context or field that consist of one or more essential characteristics. The set of characteristics that form a concept is called the concept's intension. The set of objects that are conceptualized into a concept is known as the concepts extension. A concepts intension determines a concepts extension in the sense that each object in the concepts extension must have properties that correspond with the characteristics in the concept intension.

A designation is either a name, or a symbol, or a term. Names and symbols are used to designate individual concepts. Terms are used to designate general concepts. A term is an expression that consists of one or more words that represent a general concept in a special language. For natural languages it is well known that the relationships between terms and concepts are many-to-many, due to problems of homonyms and synonyms.

In the case of homonyms, one and the same term may designate several concepts, while in the case of synonymy different terms may designate the same concept. Homonymy and synonymy can lead to ambiguity. In special languages, therefore, the objective of term-concept assignment is to ensure that a given term designates only one concept and a given concept is represented by only one term, a condition called monosemy.

The preferred method to achieve monosemy is to use intensional definitions, whereby multireferential terms are organized in hierarchical structures.

A definition is part of a so called terminological entry, which is made up of a *subject*, a *copola* and a *predicate*. The subject is the designation that is to be defined, the copola is understood to be the verb "*is*", and the predicate is the definition. Terminologists make use of intensional and extensional definitions. However, since intensional definitions are based on the relations that exist between concepts, a few words need to be said about concept systems and relations first.

Prior to the definition of terms, the concepts in a subject field are organized into a concept system. A concept system is a collection of concepts that are related by hierarchical and associative relations. In a hierarchical relation, concepts are organized into levels where the *superordinate* concept is subdivided into at least one *subordinate* concept. Subordinate concepts, at the same level, and having the same criterion of subdivision, are called *coordinate concepts*. A set of coordinate concepts are said to constitute a *dimension*. A superordinate concept can have more than one dimension, in which case the concept system is said to be *multidimensional*.

Two kinds of hierarchical relations are recognized by terminology: *generic* and *partitive*.

A generic relation exists between two concepts when the intension of the subordinate concept includes the intension of the superordinate concept, in addition to at least one delimiting characteristic. The superordinate concept in a generic relation is called a *generic concept* and the subordinate concept is called a *specific concept*.

A partitive relation exists when the superordinate concept represents a whole, while the subordinate concepts represent parts of that whole. The parts come together to form the whole. The superordinate concept in a partitive relation is called a *comprehensive concept*, and the subordinate concept is called the *partitive concept*.

In addition to hierarchical relations, there are *associative relations*. An associative relation exists when a thematic connection can be established between concepts by virtue of experience.

For example, some associative relations relate concepts with respect to their proximity in space or time, such as *container-contained*, *material-product*, *action-tool*, *action-actor*, *action-location*, *material-property*, while others may relate *cause and effect*.

Returning now to definitions, an intensional definition is always based on a either a generic or a partitive relation. That means that the definition contains a generic or comprehensive term, together with one or more characteristics that sets the term which is defined apart from other coordinate terms. What is achieved by this approach is that terms become disambiguated by the broader context of the superordinate term. The term 'Paper' for instance, has one meaning when its context is set by the superordinate term 'Publication', and another meaning in the context of 'Material'.

Intensional definitions are generally preferable to other definitions because they are more stable, useful, and informative. Consider the two examples below:

1) Planet = A celestial body that revolves around the sun in the solar system.

2) Planet = Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune, Pluto.

The first definition is intensional. If a new planet happens to be discovered in our solar system, or one of them should disappear one day, the definition is still valid. In addition, the definition provides a method to decide what counts as a planet.

The second definition is extensional. An extensional definition is defined as a list of subordinate concepts that belong to a single dimension. Extensional definitions are only valid as long as the list of subordinate concepts remains unchanged. In addition, the definition does not give any clues as to what is required for something to be a planet.

According to Bowker and Lethbridge (1994), terminology is concerned with the linguistic representation of concepts; it entails collecting, processing and presenting terms which are lexical items belonging to specialized fields. This is very similar to what Jacob (1994) understood by classification: a social process of structuring a specific knowledge domain, in order to ensure consistency and stability of meaning. In the International Standard 704 the process is described by the following steps:

1. Selecting preliminary designations and concepts by taking the subject field, the user groups and their needs into account.
2. Analyzing the intension and extension of each concept.
3. Determining the relation and position of the concepts within the concept system.
4. Formulating and evaluating definitions for the concepts based on concept relations.
5. Attributing designations to each concept.

Terminology has much to offer in an analysis of classification. The fundamental concepts fit nicely into the analysis framework, as shown in figure 2.4. The figure is not very different from that of cognitive psychology, but in contrast, the concepts are simpler and unambiguously defined.

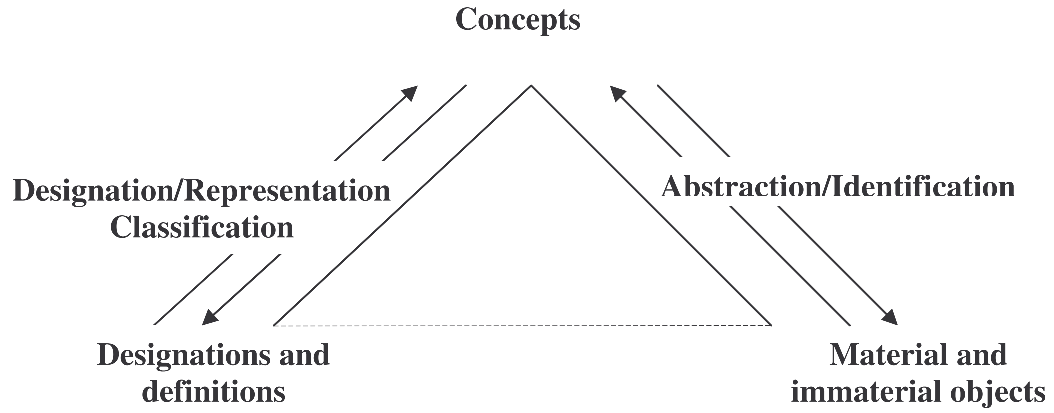

Concepts
Designation/Representation
Classification
Abstraction/Identification
Designations and definitions
Material and immaterial objects

Figure 2.4: Classification and related concepts from the perspective of terminology.

Unlike cognitive psychology, terminology has a pragmatic attitude when it comes to the nature of its fundamental concepts. In the course of producing a terminology, philosophical discussions on whether an object exists in reality are of no interest. It is taken for granted that objects exist, that they have properties, and that 'meaningful units of thought' and essential and non-essential characteristics are relative terms. Concepts are constructions, determined by subject fields, user needs, and purposes. What is meaningful or essential in one context can be meaningless or non-essential in another.

The relationships between designations and concepts are generally recognized to be many-to-many, and the objective of terminology is to produce a terminology where that relationship is reduced to one-to-one. This is where classification fits in.

The relationships between concepts and objects are many-to-one. One concept determines one extension, but the same extension may be determined by more than one concept. A classical example is 'the morning star' and 'the evening star'. Two different terms, two different concepts, but one and the same object, Venus.

Finally, the relationship between designations and objects is generally not considered, though it is a fact that more and more industry products come labeled with a name that designate which extension the product belongs to.

The fundamental concepts of terminology are displayed in figure 2.4. Since these concepts already have been treated in the text, they will not be repeated here. However, the more specific concepts have intentionally been left out, in order not to clutter the figure too much. These are concepts that are central to describe, understand and practice the process of classification, whereby concepts and designations are collected, processed and presented. For the subsequent discussion, it is important to make note of them.

Term	Meaning
Property	Not defined.
Characteristic	A generalization of similar properties.
Concept	Unit of thought
Extension	Objects viewed as a set and conceptualized into a concept.
Intension	A list of essential characteristics
Superordinate concept	A concept that is subdivided into one or more concepts.
Subordinate concept	A concept resulting from a subdivision of another concept.
Coordinate concepts	Subordinate concepts at the same level that have been divided by the same criterion.
Dimension	The set of coordinate concepts resulting from the application of the same criterion of subdivision.
Generic relation	A relation between a superordinate and a subordinate concept where the intension of the superordinate concept is included in the intension of the subordinate concept.
Partitive relation	A relation between a superordinate and a subordinate concept, where the superordinate concept represents a whole, while the subordinate concept represents a part of that whole.
Associative relation	A thematic connection that can be established between concepts by virtue of experience.
Generic concept	A superordinate concept in a generic relation.
Specific concept	A subordinate concept in a generic relation.
Comprehensive concept	A superordinate concept in a partitive relation.
Partitive concept	A subordinate concept in a partitive relation.

Table 2.7: Key concepts in terminology.

Archaeological perspectives.

Archaeology is the study of human cultures through the analysis of architecture, artifacts, and landscapes. In other words, archaeology's perspective on classification starts with the objects. For objects to be analyzed, objects must be collected, sorted, typed, labeled, named, described, catalogued and filed. These are all activities that are closely related to classification, both understood as a system and as a process. To get an intuitive idea of how important classification is to archaeology, consider the following citation from William Y. Adams, who organized the archaeological salvage campaign in Sudanese Nubia, before the building of the Aswan High Dam:

> "I had within a matter of months to organize survey and excavation programmes in an area containing literally thousands of sites, ranging in age from Palaeolithic to late medieval, with only a 50-year-old typology of graves as a starting point. It was somehow necessary for me not only to devise a strategy for sampling so large and diverse a universe, but also to create a system for cataloguing the results, and for presenting them to the public. Before I had finished I made, modified, and sometimes unmade several pottery typologies, a classification of house types, a classification of church types, and a classification of Nubian cultural periods. Most of these schemes grew from hasty and sometimes rather awkward beginnings, through successive refinements, until today they are in general use in the Nile Valley." (Adams 1988, p. 42).

Like cognitive psychology, archaeology has experienced a period of discussion about the ontological, epistemological and methodological nature of classification systems, which is known as the typological debate, (Adams, 1988; Malt, 1995; Whittaker, Caulkins, and Kamp, 1998). According to Adams (1988) the debate has been bedeviled by false and misleading dichotomies: between natural and artificial classification, essential versus instrumental types, intuitive versus rational types, induction versus deduction, lumping versus splitting, object clustering versus attribute clustering, paradigmatic versus taxonomic ordering, and empiricist versus positivist classification.

The debate is mainly a struggle between those who consider themselves active field practitioners and those who are more interested in the theoretical aspects of archaeology. From the practitioner's point of view, the great majority of types and typologies are influenced by elements from most if not all dichotomies that were debated. All types are instrumental in that they must be useful in order to be retained, and most types have evolved through a dialectic process between intuitive and rational types, inductive and deductive types, object clustering and attribute clustering, lumping and splitting.

In the course of the typological debate, people in both camps have been forced to reflect on their own practice and to address fundamental questions of classification. As a result, several papers have been published where classification and related concepts are analyzed in minute details. One such paper is Adams (1988) *Archaeological classification: theory versus practice*, which was later to be followed by a book on the same topic, by Adams and his brother Ernest Adams, (Adams and Adams, 1991). Another book which covers classification in great detail is Dunnell's book on classification in prehistory, (Dunnell, 1994).

The work of the Adamses, deals with the terminology of archaeological classification, and with the processes, purposes and practicalities of archaeological classification systems. Some of their viewpoints and definitions may be of importance to the subsequent analysis and have been extracted and collected in the tables below.

First, the Adamses distinguish between seven different senses of classification, four of which have to do with classification structures and arrangements, and three that are concerned with classification processes.

Term	**Meaning**
Classification	A kind of formal and restricted language that is made for purpose of communication, and not for sorting objects into categories. Consists of partly contrasting categories.
Typology	A kind of classification which is made specifically for the purpose of sorting objects into mutually exclusive categories.
Taxonomy	Classifications and typologies having a hierarchic feature. Where hierarchy is present, it is nearly always a secondary feature; a manipulation of the basic types after they have already been designated. Most of the time it is a way of indicating relationships between types; something that cannot be done in a basic or one-level typology because of the principle of equidistance of types.
Serialization	A linear ordering of types that have previously been created.

Table 2.8: Key concepts in archaeology related to classification structures and arrangements**.**

The next table describes three classification processes. Here, typing and sorting are two distinct versions of identification. Typing is based on categorical judgments given by 'type definitions'. Sorting is based on fuzzy judgments given by 'type descriptions'.

Term	**Meaning**
Classifying	The process of creating categories.
Typing	The process of allocating a single object to a type category based on a type definition.
Sorting	Systematic allocation of a collection of objects into type categories based on rules of thumb.

Table 2.9: Key concepts in archaeology related to classification processes.

‘Type definitions’ and ‘type descriptions’ are two of eight elements that characterize the ‘type’ concept. According to Adams, a type consists of representational, mental, and physical components, very much in accordance with the analysis framework. However, Adams adds several elements to each component:

Representational elements	Mental elements	Physical elements
Explicit type definition **Type description** **Type name** **Type label**	Type concept Type category Implicit type definition	Type members

Table 2.10: Representational, mental, and physical elements related to the notion of a typehood.

In archaeology most types are never given a formal or explicit definition. Instead they are given exhaustive descriptions and it is assumed that the definition is embodied within the description. This explains the need to distinguish between typing and sorting. If the conditions for belonging to a type are not explicitly defined, then sorting requires personal judgments that may lead to inconsistency in classification.

This lack of explicit definitions has been recognized as a severe problem by a number of archaeologists, (Beck and Jones, 1989; Whittaker, Caulkins and Kamp, 1998). Explicit definitions are needed to eliminate arbitrary judgments, and to allow an object to be identified as a member of the same type by different persons and by the same person at different times. Misclassification may have consequences for the classifier, the classification system, and sometimes also for the misclassified object. This will be further discussed in the analysis part.

Term	Meaning
Explicit type definition	A collection of characteristics that sets a type apart from all other types in a system.
Type description	A verbal and/or pictorial representation of the concept containing most if not all of the known characteristics of the type.
Type name	A name to be used in talking and writing about the type.
Type label.	A non-descriptive string or symbol used for data coding.

Table 2.11: Key concepts in archaeology related to the concretization of types (concepts).

The three mental elements in table 2.12 are very closely related, and give an intuitive feeling of a conceptualization process, where both intensional and extensional views are involved in the formulation of an implicit or tacit definition.

Term	Meaning
Type concept	A body of ideas about the nature and characteristics of a group of objects, which makes it possible for us to think of them collectively, and under a collective label.
Type category	A theoretical pigeonhole into which type members can be placed.
Implicit type definition	Every type is theoretically capable of having a type definition, but most types have only an unstated or implicit definition.

Table 2.12: Key concepts in archaeology related to mental concepts.

Finally, the physical dimension is represented by the objects. Sometimes, however, the objects classified may not be physical objects, but descriptions of physical objects. In archaeology for instance, grave typologies do not contain graves, but only plans and descriptions of graves.

An important distinction between physical objects and object descriptions is that a physical object can only be located at a single place at any one moment of time, while there can be multiple descriptions in multiple places for the same object.

With respect to object characteristics, Adams (1988) speaks of three kinds which are typical for artifacts: intrinsic, contextual and inferential. How does it look, where was it found, and what has it been used for? Contextual and inferential characteristics suggest the use of associative relations in the definition of types.

Term	Meaning
Type members	The objects that have been identified as agreeing with the description and/or definition of a particular type.

Table 2.13: Concepts in archaeology to denote objects.

Using the analysis framework, the concepts can be presented as follows:

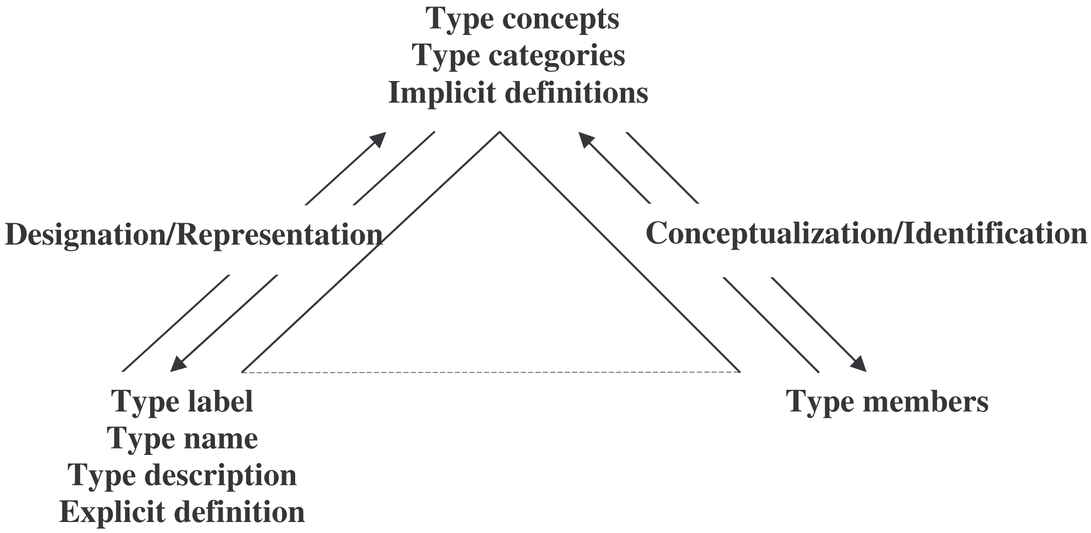


Figure 2.5: Classification and related concepts from the perspective of archaeology.

The relations between the mental and the representational elements depend on which kind of arrangement we consider. In a *classification*, which is understood to be a restricted language, Adams is willing to sacrifice precision for expressiveness, and allow the categories to be partly contrasting.

> "It is common practice, and perfectly understood for descriptive purposes, to say that the occupation of a site extended from Danubian II to Danubian IV, or that a particular component falls between Early and Middle Helladic, or that a site looks primarily Anasazi, but with a strong Mogollon admixture". (Adams, 1988, p. 44)

In a *typology*, however, which is used for sorting objects, an object can only be placed in a single type, so, formally, the types must be mutually exclusive. Since there are only one name for each type, the relations between types and names become one-to-one. However, due to the lack of explicit definitions, the risk for subjective personal judgments during the sorting process make the identification relation between the mental and the physical dimension many-to-many. As already mentioned at page 35, this may cause inconsistency in classification, leading to erroneous interpretations and conclusions, (Beck and Jones, 1989).

Conceptualization, the relation between the physical and the mental dimension, is described in the book by Adams and Adams (1991) as a continual dialectic process between the types and the objects. The process is influenced by a series of factors, such as the collection to be classified itself, the foreknowledge, the purpose of the classification, the initial and consequent analysis of types, representational decisions, and the application of the classification system

In archaeology there is always a collection of material to start with. Classifying can begin by grouping together objects or by grouping properties. How this process is carried out depends on what initial knowledge or beliefs one might have about the collection, and what purpose the classification is meant to serve. Naïve theories or background assumptions of the material to be classified may influence the selection of characteristics and types and cause soft spots in the form of unstated and untested assumptions. One indication of this is when there is no clear understanding of why particular characteristics are considered or not considered.

Purpose is another important factor that affects which types are considered relevant and which are not. Purpose is the measure of validity for any type, in the sense that there is no arrangement of types in existence that serves no purpose. Adams (1988) distinguishes between basic purposes and instrumental purposes: Basic purposes involve learning or expressing something about the classified material itself, by means of descriptive, comparative and analytic systems. Instrumental purposes involve using the classified material as a means to some other end, for instance to develop a system for storing and retrieving artifacts as a means to facilitate the subsequent description and analysis.

Yet another factor that influences classification is gestalts, i.e., clusters of objects so distinctive that they immediately suggest themselves as significant types. The first type concepts to be classified are usually intuitive types which jump at the classifiers in the form of intuitive gestalts. In the next round, the initial type concepts are analyzed for similarities and differences to disclose new types.

At some point in the classification process the type concepts must be concretized by means of terms, descriptions, pictures, diagrams or combinations of those things. Decisions must be made about which labels and terms to choose, and how much rigour and precision is required of definitions and descriptions. Precision is costly and must be balanced with the purpose for which the system is made.

Finally, classification systems change through use. Practice suggests that the classificatory process is one of continuing dialectic, or feedback, between types and objects. New material discloses new types and variants of old types that make it necessary to adapt existing patterns to new situations. Consequently types are continuously revised as long as the system is in use.

2.3.2.3 Representational perspectives on classification.

Representational perspectives include languages and modeling approaches associated with conceptual modelling in knowledge engineering and database design. It must be kept in mind that the papers have been selected because they contain one or more search terms related to classification in their title, abstract or list of keywords, so they may not be representative of the research communities to which they belong. Most papers recognize the importance of classification both as a cognitive mechanism as well as an important modeling principle. At the same time, many writers share the opinion that classification, in spite of its importance, has been generally neglected. A few citations should make this point clear:

> "Definitional language semantics is almost unheard of in the conceptual modeling tradition, where a concept description is intended to represent conditions necessary only for the extension of a class." (Bergamaschi and Sartori (1992), p 387).

> "Generally, the organization of classes/concepts into a generalization hierarchy is left entirely up to the human modeler. An interesting alternative to this practice is offered by terminological logics, where term definitions can be automatically compared to see if one is more general ("subsumes") the other." (Mylopoulos, 1998, p. 142).

> "Static constraints on subtypes are more fundamental than dynamic constraints. The only proper way to treat these is to provide formal subtype definitions and to enforce them. This point is rarely recognized in practice. It is a common misconception that the declaration is complete once subtype links (is_a connections) and exclusion and totality constraints are declared." (Halpin, 1995, p. 9).

> "The first step in constructing a conceptual model is to identify a set of fundamental concepts to describe the domain. These concepts appear in the model as classes or types. Surprisingly, there are no widely accepted rules for creating or evaluating collection of classes." (Parsons and Wand, 1997a, p. 63).

One of the earliest proposals for semantic data models was Abrial's paper on Data Semantics in 1974. The model consists of objects, categories, and binary relations between pairs of objects that belong to the categories. Although categories are used to represent sets of objects, there is no direct mentioning of classification, concepts or membership conditions. In connection with an example, it is said that:

> "The various objects are *intuitively* organized into different categories." (Abrial, 1974, p. 5).

This notion of "intuitive classification" is also used by Bubenko (1977) in one of his earlier writings:

> "Our application concerns a large set E of entities. After an examination of Q and T we 'recognize' and classify (intuitively) entities into a set C = {C_1, C_2, … , C_n} of disjoint concept classes." (Bubenko 1977, p. 65).

In a sense, the term 'Intuitive classification' seems to name a method of classification where data modeling is guided by rules on how to *find* existing classes. According to Parsons and Wand (1997a):

> "The object-oriented and semantic data modeling literature also offers advice on identifying classes. However, this advice is usually very general, such as identifying tangible things, roles and events, and is not much help in determining whether a selected set of classes is appropriate." (Parsons and Wand, 1997a, p. 64)

This idea of identifying classes is also supported by Artz (1997) who claims that most modeling methods implicitly assume that object classes do exist in the world, waiting to be discovered by the data modeler. Artz mentions three reasons for this.

First, many database systems are replacing older systems in which classes are already defined. Second, in common speech, we frequently use general terms to refer to individual objects suggesting that many objects belong to a single class. Third, some classes, with which we have first hand physical experience, seem so natural that it is hard to see what other way the individual in that class could be identified. This last reason appears to have very much in common with the so called 'intuitive types' in archaeology, which "jump at the classifiers in the form of intuitive gestalts".

Something that speaks against' intuitive classification' is the fact that several prominent writers in the field take a predicate view on classification. In Chen's famous paper from 1976, a test-predicate is explicitly mentioned for both entity sets and value sets:

> "Entities are classified into different entity sets such as EMPLOYEE, PROJECT and DEPARTMENT. There is a test predicate associated with each entity set to test whether an entity belongs to it. For example, if we know an entity is in the entity set EMPLOYEE, then we know that it has the properties common to the other entities in the entity set EMPLOYEE. Among these properties is the afore mentioned test predicate". (Chen, 1976, p. 11).
>
> Values are classified into value-sets. There is a predicate associated with each value set to test whether a value belongs to it. (Chen, 1976, p.12).

In addition to the explicit mentioning of a test predicate, Chen also makes a very important distinction that is generally not recognized in conceptual modeling: a) that entities in an entity set have properties common to the other entities in the entity set, and b) that, among these properties, there is also a test predicate. Usually, descriptions of entity sets and similar constructs contain a), but not b). The distinction that Chen makes, suggests a distinction between definitional properties and descriptive properties. While the definitional properties, which constitute the test predicate, necessarily must be invariant and the same for each and every entity in the entity-set, the descriptive properties will normally be variant and dissimilar for most entities. Take for instance a class of male students. While the defining predicate for the class requires every student to be of the same sex = 'male', properties like ssn, name, date of birth, address, phone, etc., will differ for each student, and some of the properties will even change over time. More will be said about this in the subsequent discussion.

This distinction is not directly noticed by Smith and Smith, (1977), who wrote a paper where they combined research on aggregation from the database area with research on generalization from the Artificial Intelligence area. The authors stress the importance of classification in designing database models, but use the term *generalization* instead of classification:

> "We will use the term generalization in the following way: A generalization is an abstraction which enables a class of individual objects to be thought of generically as a single named object. Generalization is perhaps the most important mechanism we have for conceptualizing the real world… In designing a database to model the real world, it is essential that the database schema have the capability for explicitly representing generalizations." (Smith and Smith, 1977, p. 107).

Although they do not mention any distinction between defining and descriptive properties, they clearly recognize a distinction between individual attributes and class attributes. In addition, they describe and demonstrate a method for representing a generic hierarchy as a hierarchy of relations. Their method requires that the immediate subclasses of any class be partitioned into mutually exclusive classes. A class may be partitioned into as many groups of mutually exclusive classes, as there are criteria for subdivision. Each group of mutually exclusive classes is called a 'cluster', and each criterium of subdivision is called an 'image domain'. An image domain is a domain that can be used to divide a class into as many subclasses as there are domain values. The domain name reflects the criterium for subdivision, while each of the domain values corresponds to a subclass.

In relational terms, this means that for a relation to be partitioned into subrelations, it must contain one attribute whose name reflects the criterium for subdivision, and whose values name the respective subrelations for that partition.

The ideas of generalization hierarchies presented by Smith and Smith (1977) are included and further detailed in a paper by Codd (1979) where he proposes extensions to the relational model in order to capture more of the semantics in a database. Codd distinguishes between two extensional aspects of generalization: *instantiation* and *subtype*.

Both are forms of specialization, and their reverses are forms of generalization.

The extensional counterpart of instantiation is *set membership*, while that of subtype is *set inclusion*. In a paper by Odell and Ramackers (1997), the two extensional aspects were later formalized and termed *classification* and *generalization*, respectively.

Codd also recognizes that entities may belong to (or be described by) several types, and that there is a need for unique and permanent identifiers to keep track of those entities for which there may be several descriptions. Here, Codd identifies the descriptive aspect of types, but he fails to recognize any definitional aspects. According to Codd, classification, or generalization by membership, as he calls it, is taken care of by E-Relations. An E-Relation is a unary relation where entity types are represented by a name only. The general idea is that the predicate or membership condition should be reflected by the name. This is clearly an example of intuitive classification.

On the other hand, when it comes to generalization by inclusion, Codd distinguishes between unconditional and conditional generalization. Both kinds are represented by a triple relation (SUB:m, SUP:n, PER:p), where m represents the subtype, n, its supertype, and p the predicate. The strange thing here is that predicates are explicitly used to control membership in subtypes, but no predicates are stated for the supertypes, or types which do not participate in generalization hierarchies.

Another well known paper from this period is Hammer and McLeod’s (1981) paper “Database Description with SDM” where they propose a model to formally specify the meaning of a database. With respect to classification, they propose that a class should have:

> “An optional textual *class description* describes the meaning and contents of the class. A class description should be used to describe the specific nature of the entities that constitute the class and to indicate their significance and role in the application environment.” (Hammer and McLeod 1981, p. 407).

Although the ‘class description’ as they call it is optional, it is a very clear request for adding a membership condition to a class. In addition, and very similar to Codd, they propose predicate-defined subclasses:

> “In SDM, a subclass S is defined by specifying a class C and a predicate P on the members of C; S consists of just those members of C that satisfy P”. (Hammer and McLeod 1981, p. 408).

The authors distinguish between different kinds of predicates and subclasses, such as attribute-controlled subclasses, user-controllable subclasses, set-operator-defined subclasses and existence subclasses. Among these, user-controllable subclasses are special in the sense that membership in the subclass is directly and explicitly controlled by the database users. The reason for allowing user-controllable subclasses is, according to the authors, that in some cases, the membership condition is too complex to be recorded in the database schema. This may also explain why the ‘class description’ was proposed as an optional entry. A similar reflection is done by Norrie (2000). Referring to philosophers, linguists and psychologists, he holds that concepts cannot always be defined by necessary and sufficient conditions. Therefore he suggests that classification may be left to the entities themselves when filling in forms, or left to the end-users when recording or updating data.

This mix of statements from both pioneers as well as more recent writers in the field, indicate that classification is primarily based on intuition, though other approaches to classification are also noticed to a certain degree.

Among the remaining papers, one relatively detailed account of classification is given by Odell and Ramackers (1997) who make an attempt to formalize the key concepts of object-oriented analysis. In the paper they define concepts and emphasize important distinctions between concepts which are fundamental to an understanding of classification in the context of conceptual modeling. The key concepts are summarized below.

Terms	Conceptualization	Analysis	Design
Concept	Idea or notion that we apply to classify things around us.	A type definition that represents a concept.	Class definition that implements a type
Intension	The meaning of a concept. An identification rule.	An intensional definition of a type.	Identification method of a class
Extension	One or more objects to which a concept applies.	One or more objects to which a type applies.	One or more OO-constructions that implement a type.
Type	Idea or notion that we apply to classify things around us.	A type definition that represents a concept.	One or more OO-constructions that implement a type.
Class	OO-implementation(s) of a concept.	A type definition that represents a concept.	One or more OO-constructions that implement a type.
Object	Anything to which a concept applies.	An instance of a type	An instance of a class.
Classification relationship	A relationship between an object and a concept that applies to it.	An assertion that a given object is an instance of a given type.	A pointer or reference from the object to the class that constructed it.
Generalization/specialization relationship	A relationship between two concepts.	A subsumption relationship between two type definitions.	A subset relationship between two class extensions.
Spesialization or subtype	The left side of a gen/spec relation.	Any type A, whose definition contains the definition of another type S.	Any class A, whose members are included in another class S.
Generalization or supertype	The right side of a gen/spec relation.	Any type S, whose definition is contained in the definition of at least one other type A	Any class S, whose members also belong to one or more subsets.

Table 2.14: Key concepts related to object-oriented analysis and design.

The important thing to notice here is the authors' claim that concepts mean different things in different contexts such as analysis and design. Because of this, a concept for instance, is called a type in analysis and a class in design. Types define a problem, while classes represent a solution to the problem. Therefore, analysts are concerned with terminology and definitions in order to define the problem, while designers are concerned with efficient storage structures and inheritance in order to design an efficient solution.

The distinction between intensional and extensional aspects has also been focused on by AI's research on knowledge based systems. Research in the 80's concentrated on so called classification-based languages such as KL-ONE and CLASSIC. A classification-based knowledge representation language includes two languages: a terminological language and an assertion language. A terminological language is used to define classes of individuals in a particular domain.

An assertional language is used to state constraints or facts that apply to a particular domain. (Mac Gregor, 1991; Mylopoulos, 1998; Woods, 1991). The key concepts that are used to characterize a terminological language are summarized below:

Term	Meaning
Concept	An abstract conceptual entity that is characterized by an intension and an extension.
Concept definition	A concept definition states a necessary and sufficient condition for membership in the extension of a concept.
Taxonomy	An organization of concepts based on a subsumption relationship that relates pairs of concepts.
Subsumption	A concept C subsumes a concept D if any individual satisfying the definition for D necessarily satisfies the definition of C. Thus, if C subsumes D, then the extension of C is a superset of the extension of D.
Classification	The process of inserting a new concept into a taxonomy of concepts so that the more general concepts are positioned above it, while less general ones are positioned below it.
Type	The set of concepts that an individual in a knowledge base belongs to is called the type of that individual.

Table 2.15: Key concepts in AI related to knowledge representation.

The terminological aspect is also very central in more recent AI-research on ontology building and use. According to Guarino (1997), one of the main motivations for this research is the possibility of knowledge sharing and reuse across different applications. To achieve this, the applications must commit to the definitions in a common ontology. An ontology in this respect defines the terminology of a domain of knowledge, (Waterson and Preece, 1999), and the commitment, which is known as 'ontological commitment' is defined as an agreement to use a shared vocabulary in a coherent and consistent manner (Gruber, 1995). This has much in common with the goals of conceptual modeling, though the focus on vocabulary and terminology is less emphasized.

In closing this section, the following can be said with reference to the analysis framework: Among the selected papers, there are at least two major views on classification. One view discover existing classes by looking at the universe of discourse, the other view constructs classes by looking at people's knowledge of the universe of discourse.

According to the first view, classification relies heavily on intuition. As an example, see the last paragraph in the quote at page 1. Intuitive classification is a matter of identifying the types that exist in the universe of discourse. There are no formal rules, except for some simple heuristics, such as looking for people, things, roles, interactions, and places.

As soon as the types are found, the relevant objects are expected to fall into their respective types automatically, so identification of members is not an issue. With respect to the analysis framework, classification can be viewed as a direct relation between a type and its associated objects, without any further references to mental concepts and terminology.

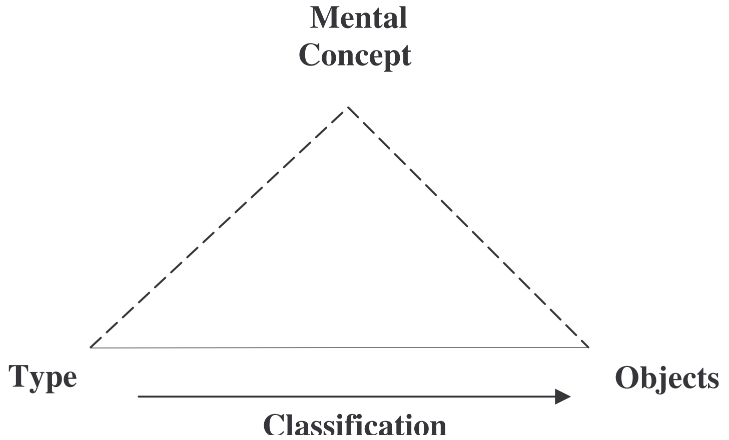


Figure 2.6: Classification and related concepts from the perspective of intuitive classification.

The other view is more complex, and involves the interplay among cognitive, linguistic, and ontological elements. Here, classification has very much in common with terminology, where mental concepts are concretized and formally defined before they are arranged in a system of concepts.

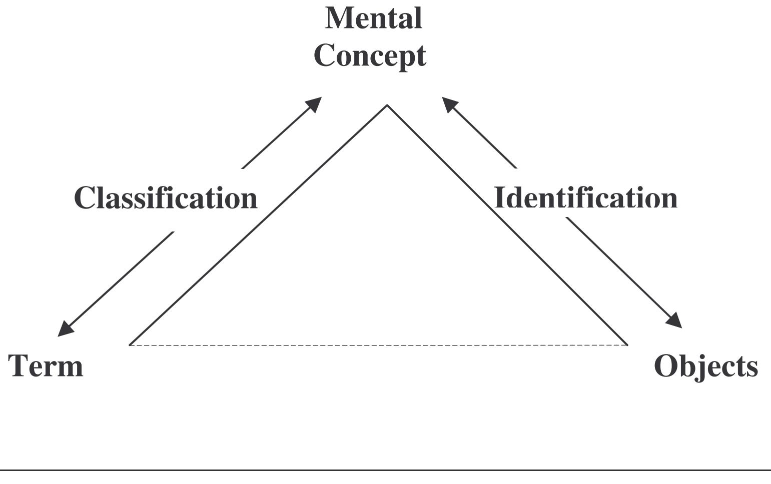


Figure 2.7: Classification and related concepts from a constructivist/pragmatist perspective.

In accordance with this view, a mental concept is understood as an abstract conceptual entity that is characterized by an intension and an extension. The concepts intension is represented by a predicate, which, in turn, determines the concepts extension, i.e., those objects that satisfy the predicate. We may say that knowing the intension is a prerequisite to knowing the extension:

> If we understand the intension, we are acquainted with the extension – even if we have never seen an individual that belongs to it. (For example, we may possess the concept of Unicorn without ever seeing an instance of one.) However, the converse is not necessarily true. If we understand the extension, we do not necessarily know the intension. (Odell and Ramackers, 1997, p. 2).

In order for concepts to be shared and communicated, concepts must be defined. This process of defining concepts and finding their correct position in a hierarchy of definitions is called *classification*. Classification produces a terminology that is suited to name the concepts that make up the universe of discourse.

Similar to the relation between intension and extension, one may say that classification is a prerequisite to identification. One cannot identify something by a name alone, unless one knows how to apply the name correctly. Similarly defining concepts is a prerequisite to modeling concepts, since terms cannot be related to form a concept system unless one knows their definitions. Finally, we may say that identification is a prerequisite to description. It goes without saying that nothing can be described before it has been identified. I cannot describe something as a flower unless I have it before me and it has been recognized as a flower.

All distinctions referred to above will be further discussed in the next section.

2.4 Discussion

As indicated by the findings in the preceding section, classification cannot be understood in isolation, but only in relation to a number of other concepts. Table 2.16 contains a list of general ideas that form the necessary basis for any discussion of classification. Therefore these ideas are first defined and discussed, before they are used to clarify and discuss the concept of classification.

General ideas	Cognitive Psychology	Terminology	Archaeology	Knowledge Representation.	Database Design
Universe of discourse	Contexts, theories, goals, needs.	Subject field, purpose, user needs.	Collection, purpose.	UoD, Domain.	UoD, mental models.
Concept	Concept	Concept	Type concept	Concept	Concept, type
Intension	Categorization rule, mental model	Essential characteristics	Explicit type definition, Implicit type definition	Membership condition	Identification rule, Test predicate
Extension	Category	Set of objects	Type members	Set of instances	Set of objects
Object	Category member, real object, artificial object	Anything to which a concept applies	Phys. object, description	Instances of concepts, individuals	Instance of type or class
Property	Property, feature, characteristic	Property, characteristic	Intrinsic, Contextual, Inferencial	Concept	Attribute, domain, value
Arrangement	Taxonomy, partonomy, mental models	Concept systems, taxonomies partonomies associations	Classification Typology Taxonomy Seriation	Taxonomy	Type systems, taxonomies, partonomies, associations

Table 2.16: General ideas that are commonly related to classification.

2.4.1 Definitions.

A formal context: A collection of concepts and relationships that *may be explicitly defined and agreed upon* for the purpose of serving a certain discourse (Langer, 1967).

The idea of a formal context is taken from symbolic logic, and introduced here to distinguish the psychological context of categorization from the social context of classification. This is directly related to Jacob's interpretation of ‘categorization’ and ‘classification’. While categorization is a cognitive process that can operate on private and subjective concepts, classification is seen as a social process that operates on ‘public’ concepts that can be crisply defined.

The Universe of Discourse*, *(UoD): The total collection of *concept definitions* belonging to a formal context, Langer (1967).
Although the term ‘Universe of Discourse’ is not widely used, some kind of demarcation is recognized by all disciplines, in terms of contexts, purposes, goals, collections, user needs, and so on. With direct reference to conceptual modeling, the universe of discourse can be understood as the end result of classification, i.e., *the total collection of concept definitions that are required to serve the purpose of the information system.*
If one regards the UoD as the input to the modeling process, this suggests that classification, as a process, should be performed prior to the modeling process.

Concept: A ‘concept’ is the general idea of a mental construct, and thus a subjective and private construct, which is variously termed ‘idea’, ‘conception’, ‘conceptualization’, ‘conceptual unit of meaning’, or ‘unit of thought’. According to the view held by most of the surveyed disciplines, a concept will be defined as *a mental representation of a condition that 1) is used to identify objects to which the concept applies, and 2) any object must satisfy in order to be considered an instance of the concept.*
The definition emphasizes only two functions that the concept is meant to serve, and that is first, to identify objects, and second, to allow us to think of a group of objects collectively, and under a collective label. The notion of identity here is not one of individual identity, but rather of what Adams (1988) calls typological identity. Typological identity is a way of recognizing similarity among objects, as opposed to individual identity, whose purpose is to distinguish between objects.
It may well be that concepts serve many more cognitive functions, as cognitive psychologists claim, but in order to serve classification in the context of conceptual modeling, it is sufficient to regard a concept as consisting of a *set of references* to properties that any object must possess to be considered an instance of the concept. As conceptual units of thought, concepts themselves have no properties. The concept of a horse does not smell like a horse, nor does it look like a horse in any respect. Hence, concepts cannot be described, because nothing can be said of them. Concepts can only be defined. However, it does make sense to describe individual horses, once they have been identified as ‘horses’, or even to describe the whole group of horses so identified. This distinction between definition and description is also reflected in the distinction between the intension and the extension of a concept. Definitions depend on the intension of a concept, while descriptions depend on the extension.

Intension: The intension of a concept is understood as the condition that the concept stands for. Formally the condition is referred to as a set of references to properties that are *singly necessary* and *jointly sufficient* for an object to fall under the concept. For a property to be singly necessary, every object must have it. For a set of properties to be jointly sufficient, every object having that set of properties must be an instance of the concept.

In this respect a concept definition states a set of properties that must be invariant and the same for all objects that satisfy the condition.

Extension: The extension of a concept is understood as a logical class of all and only those objects that satisfy the concept's intension. This means that all objects that belong to the same class have a set of definitional properties that are invariant and the same to all of them. However, in addition to these invariant properties, each object has a set of descriptive properties that are variant, and often different from at least some of the other objects in the class.

By means of identification, a concept's intension is used to assign objects to the concept's extension. However, different conditions may call for different identification procedures. Consider the conditions below regarding how to decide whether a person is a female or a male athlete:

1. If a person looks and acts like a male athlete, identify the person as a male athlete, otherwise as a female athlete.
2. Ask the persons for their sex, and identify them according to their answers.
3. Take a urine sample and run a biometrical test according to procedure A1.

In 1) one has to rely on the personal judgment of the classifier, in 2) one has to rely on the classified, and in 3) one needs special procedures and extra, technological equipment to decide on the identification.

This example, which has been developed from ideas in Paul Starr's paper "Social categories and Claims in the Liberal State", (Starr, 1992), reveals at least four questions that must be considered for each and every concept to be defined: Why has exactly this condition been chosen? How do we apply the condition? Are there any consequences of misclassification? Who said so? These questions will be further treated under the discussion of 'classification'.

Object: A simple definition of 'object' is to say that an object is an *instance* of a concept, type, or class, (Odell and Ramackers, 1997). To say that an object is an instance of a concept (in KR terminology), or type, (in Odell and Ramacker's terminology), is similar to saying, as the terminologists do, that an object is *anything to which a concept applies*. In these respects, an object may be understood as a physical object that exists in time and space, or something abstract that can only be conceived of, such as the idea of a triangle.
However, different purposes may cause an object to be variously *described*, so that a single *instance of a concept* may be associated with several *descriptions*. As mentioned by Adams (1988), descriptions of objects may themselves be considered as objects. Still there is a difference between the two: while a physical object can only be one, there may be multiple descriptions of the same object. This observation calls for a distinction between *instances of concepts* and *instances of a types*: an instance of a concept is always a physical or abstract object, but an instance of a type is always a description of an object.
Related to classification and conceptual modeling, the universe of discourse contains *concept definitions* that specify the defining properties of objects, while the conceptual model contains *type definitions* that specify the descriptive properties of objects.

Property: Objects are generally said to possess properties, and a property is said to represent an aspect of an object. With respect to classification, properties are known as the *principles of classification*, (Grossmann, 1992). Objects that possess the same properties are said to constitute a class. Any property may serve as a principle of classification, and the classes that result depend on which properties that have been chosen to do the work.
While some properties, in a certain context, may be considered relevant for classification purposes, other properties may be considered relevant for identification or description purposes. Hence, one may distinguish between defining properties, identifying properties and describing properties:

- A ***defining property*** is a property that represents an objects typological identity, i.e., an invariant aspect that is exactly the same to all objects in a class.
- An ***identifying property*** is a property that represents an objects individual identity, i.e., a variant aspect that is distinct and unique to each object.
- A ***describing property*** is a property that represents a variant and descriptive aspect of each single object in a class, or of the class as a whole.

Defining and describing properties are relative terms. Properties that serve as defining properties in one context, may be irrelevant, or serve as describing properties in another context and visa versa.

Arrangement: The term ‘classification’ can be associated with different kinds of arrangements, all of which are created to serve one purpose or another. An arrangement consists of types that may be *mutually exclusive* or *overlapping*, *unordered* or *ordered*. Mutually exclusive types are typically required when the purpose is to *sort* objects. One example of such an arrangement may be a set of bins to store different kinds of objects. Since an object can only be placed in one bin at the time, the types must necessarily be mutually exclusive.

Overlapping types may be preferred when the purpose is to *describe* objects. One example may be a purely analytical arrangement of idealized types, such as Max Weber’s four types of social action. When this kind of arrangement is used to describe instances of observed, human actions, then some observations may be best described by multiple types. Arrangements, that are used to facilitate communication and description, are called ‘terminologies’, ‘special languages’, or ‘vocabularies’.

An unordered arrangement of types is variously called a ‘paradigm’, ‘typology’ or ‘faceted classification’. Such arrangements are normally considered to contain a set of basic types, which may later be manipulated into various orders.

An ordered arrangement of types is called a ‘serialization’, ‘taxonomy’, or ‘partonomy’. A ‘serialization’ refers to an order where a set of basic types are linearly ordered. A ‘taxonomy’ refers to a hierarchic order whereby types are related by generic relations, and a ‘partonomy’ refers to a hierarchic order whereby types are related by partitive relations.

There are, according to Adams (1988), no arrangements that serve no purpose. Hence, purpose is the measure of validity of any arrangement.

Classification: In the context of conceptual modeling, the term ‘classification’ can be used to mean several things: **1)** a process to define the concepts and relationships that constitutes the universe of discourse; **2)** the set of terms and definitions that result from 1); and **3)** a process to identify and verify the typological identity of objects.

Classification as a process of defining the universe of discourse: The general idea is simply to start by defining the concepts and relationships that make up the universe of discourse. After all, it is a reasonable requirement that one is let to know what there is to be talked about. This is not a mandatory step in any methodology as far as I know, but it could be a good idea to maintain an "inventory" as the process of conceptual modeling advances. In this sense, 'classification' is understood as *a social process between users and analysts, where concepts are concretized and reconciled into a common vocabulary*. For concepts to be concretized they must be defined, and the preferred kind of definition is definition by intension.

Ideally, concepts should always be intensionally defined, but this may sometimes be an unrealistic or unnecessary requirement. Anyone making a personal database application to keep track of a personal cd-collection would probably not see any point in spending time on concept definitions.

When should intensional definitions be required, and when can this requirement be relaxed upon? One possible answer to this is indirectly given by Paul Starr (1992):

> "Classifications have consequences. Some cause damage; some advantage. That is, above all other reasons, why people fight over them". (Paul Starr, 1992, p. 161)

Classification, or misclassification, may have serious consequences, sometimes for the classifier, sometimes for the classified, and sometimes for the system as a whole. As an example, consider the consequences of misclassifying edible and poisonous mushrooms. The more serious the consequences, the more important it becomes to control and guard against misclassification.

This suggests that during the classification process, the *consequences* for misclassification should be analyzed and documented for each concept.

Such an analysis should take both the membership condition, as well as the consequences of misclassification into consideration. Typical questions that need to be answered are: Why exactly this definition? What may the consequences of misclassification be? Who or what is the source of this requirement? How is the membership condition to be controlled?

The results from this analysis may help to decide how much precision is required of a definition and how the definition is to be controlled. In extreme cases, it can result in an identification procedure, or algorithm that needs to be implemented during design.

Classification as a set of terms and definitions. The end result of the classification process is a set of terms and definitions that constitute the universe of discourse. In terminological terms, the end result is sometimes called a ‘terminology’, and sometimes a ‘controlled vocabulary’, or ‘restricted language’. Here, the end result is called the universe of discourse, meaning a special language where each term designates a single concept, and each concept is represented by a single term.

Because most concepts are intensionally defined, the concepts are arranged into hierarchies of superordinate, and subordinate concepts. In addition, it is common practice in terminology to relate concepts by partitive and associative relations. Since these are the same kind of relations that are used in conceptual modeling, it is important to consider the difference between the universe of discourse and a conceptual model. The main difference is that a universe of discourse is concerned with *terms* and *definitions*, while a conceptual model is concerned with *types* and *descriptions*. While definitions are used to identify objects, the types are used to specify which properties the objects will be described by. It sounds quite reasonable that definition necessarily must precede description, for without a definition there is no way to identify what is to be described. Also, for objects to be identified as instances of a concept, they must have the necessary defining properties. *This means, that classification, in the context of conceptual modeling, should be executed prior to the conceptual modeling process*. Section 1.1 contains some further arguments that support this claim.

Classification as a process to identify and verify the typological identity of objects: Conceptually, this sense of ‘classification’ is an act of judgment, where objects are identified as instances of a concept. Sometimes the term ‘*identification’* is used to distinguish it from other senses of classification. Technically, in a database setting, identification can be understood as a process that is carried out whenever data are entered into the database.

In other words, this sense of classification is related to operational aspects of an application. As a process it is either user-controlled or controlled by the application. In both cases, procedures are needed to match the rigor and precision required by the membership condition. In addition, decisions must be made with respect to their implementation, either as user-related, administrative routines, or as methods to be automatically triggered and run during insert-operations. Most often, the procedures can be expected to develop gradually from problem statements during the initial classification phase, via a gradual refinement of algorithms during conceptual and logical data modeling, until its implementation in the user-manual or in a program.

2.4.2 The resulting conceptual framework

The definitions presented in this section are products of the current concept analysis. Taken together, the definitions constitute a conceptual framework which meets the overall purpose of the current study. To get a more condensed and materialized view of the framework, table 3.2, on page 63, shows how the definitions from section 2.4.1 are used to guide the data collection in chapter 3.

2.5 Implications and hypotheses.

As the results of this study show, classification is closely bound up with the notion of defining properties. Defining properties are recognized by all disciplines that have been surveyed, and have been explicitly described in Chen's (1976) paper on the Entity-Relationship Model. In spite of this, defining properties seem to be neglected during conceptual modelling, except when it comes to so-called 'attribute-controlled sub-classes', (Smith and Smith, 1977; Hammer and McLeod, 1981).

What would the implications be if classification were introduced as an overall requirement to conceptual modelling?

2.5.1 Shared understanding of basic concepts.

Classification is generally recognized as a fundamental abstraction mechanism for conceptual modelling, and software engineering, (Booch, 1991; Odell and Ramackers, 1997; and Mylopoulos, 1998). Yet, in spite of its claimed importance, the discipline seems to lack a unified account of what classification is, how it is performed, and why it should be performed. These assumptions are explored in chapter 3, where text-books in conceptual modelling are reviewed in order to establish exactly what classification means, compared with the views developed in this chapter.

2.5.2 Completeness and logical consistency.

A classification in the form of a vocabulary may be evaluated for its completeness and logical consistency. The vocabulary is complete when it includes all the concepts mentioned in the information requirements, (Moody and Shanks, 1998), and it is logically consistent when all concepts are correctly defined by intensional definitions. The hierarchical structure that results from intensional definitions may easily be checked for its logical consistency. This aspect is further pursued in chapter 4.

2.5.3 Basis for modeling decisions.

The logical and hierarchical structures that result from the classification process may guide the naming, selection and justification of entity types, relationship types and roles in the conceptual and logical models. Since the conceptual level is mainly concerned with the intensional aspects of concepts, design decisions at this level will be motivated by intents to make the model as simple as possible, yet rich enough to convey the meaning of the concepts.

At the logical level, which is mainly concerned with extensional aspects, conceptual structures may be inflated or conflated to support dynamic classification, multiple classification, and inheritance. Dynamic classification refers to the ability to change the classification of an object (Odell, 1998). At one moment an object may be a student. At a later moment the same object is declassified as student and classified as an employee. Multiple classification means that an object may be an instance of more than one object type (Odell, 1998). A student assistant for example, may be an instance of both student and employee.
At both the conceptual and the logical levels, the vocabulary will provide a framework for discussing the design decisions that are made. (See section 4.4.5, on page 113 for an example). These aspects are further pursued in chapter 4.

2.5.4 Organizational, administrative and technical implications.

Classification draws attention to the *consequences* of misclassification, and to the *organizational*, *administrative*, and/or *technical* measures that need to be taken to avoid unwanted consequences. A failure to define the concepts correctly may result in database integrity problems or incorrect information to unwary users (Artz,1997). For example, what does it mean to be a member of a political party? Does one count as a member if one agrees to be counted as a member, or does one have to pay a membership fee? What happens if the membership fee is not paid for? A user asking for the member status will receive a number in response. What this number means depends on how the concept of *member* is defined. Political parties receive financial aid based on their membership reports. The consequences of reporting incorrect numbers may be that the party looses its credibility, staff has to resign, money aid has to be returned, perhaps along with a fine, and so on.
To avoid this from happening, a membership condition needs to be stated, someone need to be appointed as responsible for the member file, administrative routines must be settled regarding the registrations, updates and deletions of members, along with technical issues to implement the routines. Consequences of misclassification are further pursued in chapter 4.

2.5.5 Data integrity

Membership conditions can be used to control that objects that enter a class really belong there. If users are unaware the membership conditions for a class, incorrect instances may be recorded. Hence, for class-based applications, membership conditions may be formalized and controlled by the application. At the conceptual level, the membership condition can be expressed in natural language, as it appears in the vocabulary, but in addition it may need to be formalized and represented by a defining property.

At the logical level, membership conditions may be operationalized and represented by an algorithm for the actual checking that must be done. At the physical level, the algorithm may be implemented by means of constraints, triggers, procedures or methods, depending on the chosen DBMS.

As an example, consider a Student entity type. The predicate associated with Student is that the *term fee must have been paid*.

In the vocabulary, the definition becomes:

> ***Student****: A student is a person who has paid the term fee.*

In the conceptual model, the predicate is transformed into a defining attribute and necessary, administrative tasks associated with the identification process are described.

> *A date attribute is chosen as the defining attribute. To be classified as a student, a bank receipt is needed that confirm the payment, and the transaction date on the receipt must be entered into the defining attribute.*

In the logical data model the identification procedure may be expressed by an algorithm:

> *On Insert Into Student check feeDate is not NULL.*

In the physical data model, the algorithm may be implemented as a simple NOT NULL constraint.

> *feeDate Date NOT NULL,*

The example demonstrates how a membership condition can be transformed into an identification procedure that takes care of the database integrity. In general, a membership condition at one level may be considered as a problem statement, for which there may be many solutions at the next level. Consequently there is a one-to-many relationship from a statement in a vocabulary to its representation in the conceptual model, from the conceptual model to the logical data model, and from the logical data model to the physical data model. This is further elaborated in chapter 4.

2.5.6 Validation and interpretation of conceptual models.

Classification represents an addition to the semantics represented by a conceptual model. One example may be that membership conditions may help to explain the participation and multiplicity constrains on relationships. Without knowing the membership conditions of the entity types involved in a relationship, it may not be possible to decide whether the constraints displayed in the model are correct. Consider the following example:

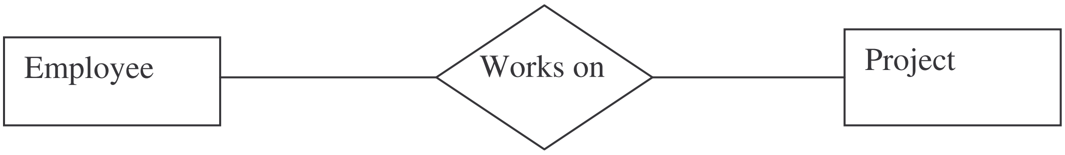
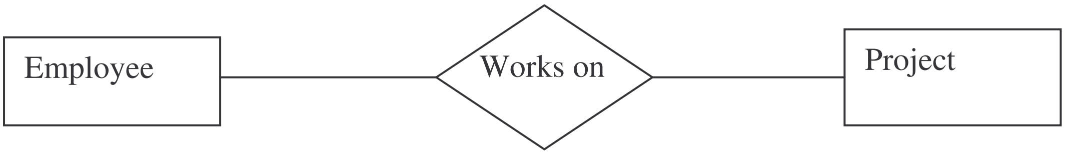


Figure 2.8: Example of a binary relationship.

If one is asked to assess the participation and multiplicity constraints of the *Works on* relationship between *Employee* and *Project*, we can only rely on common sense and general knowledge about similar cases. In a case like this, a reasonable interpretation would be that employees *may* work on *one or more* projects, and that a project *may* have *one or more* employees at work.

Now, repeat the assessment, by using the information from the membership conditions that are added below each entity type.



An employee is a person who
works on at least one project

A project is a complex work
task which involves the joint
effort of several employees.

Figure 2.9: Example binary relationship with membership conditions included for the entity types.

By taking the membership conditions into consideration, the most reasonable interpretation would be that employees *must* work on *one or more* projects, and a project *must* have *one or more* employees at work.

The membership conditions have changed the participation constraint from optional to mandatory, and confirm our assumption about the cardinality ratio as one or more. The point here is to show that knowledge of membership conditions may have a positive effect on the interpretation of the relationship, and on the confidence one may have in the interpretation. Knowing the membership conditions may be of great value to auditors, users, and systems analysts when they have to validate or interpret existing models. How membership conditions may affect interpretation tasks is pursued in chapter 5.

2.5.7 Data integration.

When attempts are made to integrate two or more separate applications, problems related to homonymous and synonymous types are easily confused with differences in attributes. We tend to assume that two types are homonymous if they have the same name, but differ in their attributes, and that they are synonymous if they differ in names but have similar attributes. However, if one introduces typological and individual identity as well, the picture of homonymous and synonymous types becomes a lot more shaded:

	Same attributes				**Different attributes**			
	Same MC		**Different MC**		**Same MC**		**Different MC**	
	Same IID	**Diff IID**	**Same IID**	**Diff IID**	**Same IID**	**Diff IID**	**Same IID**	**Diff IID**
Same Typename	Keep the most current copy	Consider to run a Union operation	Keep apart?	Keep apart?	Consider to run a Join operation	Keep apart?	Keep apart?	Keep apart?
Different Typename	Keep the most current copy	Consider to run a Union operation	Keep apart?	Keep apart?	Consider to run a Join operation	Keep apart?	Keep apart?	Keep apart?

Table 2.17: A tentative framework to analyze schema integration problems.

The framework in table 2.17 can be further extended by considering attribute names, data types and the number of attributes in each type. However, its current format is sufficiently detailed to show that membership conditions and individual identities are more important to consider than name-differences. This suggests that classification may be important, not only to data modelling but also to schema integration.

2.6 Conclusion

Classification discloses a distinction between the definition of concepts and the description of objects, which suggests a distinction between *concepts/terms* on one hand, and *types* (as logical data structures) on the other. The fact that concepts/terms are represented by vocabularies, and types by conceptual models, leads to the conclusion that classification may be considered as a prerequisite for conceptual modelling, and that the modelling process should be (logically) divided into two separate tasks: *the classification task*, which is concerned with the definition of concepts, and the *modelling task*, which is concerned with representational and descriptive aspects.

Methodological aspects concerning classification and modelling is fully covered in chapter 4. Before that, chapter 3 contains a content analysis of text books on conceptual modelling, to establish the status of classification in classical and current text.

3.0 Classification concepts in conceptual modelling

3.1 Introduction.

The following study reviews how, and to what extent, classification and related notions such as concept, class, object and property are defined and used in text books on data models and data modelling. The assumptions are:

a) that the notion of classification is not sufficiently attended to by the data modelling community, and that

b) lack of attention causes related notions, such as concept, class, object and property to be missing, unclear, ambiguous and/or inconsistent in their definitions.

Accordingly, the purpose of this review is to establish how classification and related notions are presented in current text-books, by comparing relevant terms and their definitions from each text book with the definitions developed in chapter 2.

3.2 Method.

3.2.1 Identification of text books.

The search process was based on text books that were available for searching and loan ordering via BIBSYS. For further details about the BIBSYS holding database, see chapter 2. The search process was run against BIBSYS in two separate rounds.

In the first round, the search process was based on free text searches, where terms from table 2.2 and 2.3 were used in combination with terms like “Practical approach?”, “Introduction”, “Advances”, “Principle?”, “Standard?”, “Fundamental?”, “Guide?”, “Modern”, and “Method?”.

In the second round, the search process was based on the reference lists of review papers and bibliographies. A list of internationally well recognized scholars was compiled, and separate searches were made for text books written by the authors in the list.

Booch, G
Brodie, ML
Bubenko, JA
Chen, P
Coad, P
Codd, EF
Date, CJ
Elmasri, R
Embley, DW
Jacobson, I
Kent, W
Mellor, SJ
Mylopoulos, J
Navathe, SB
Martin, J
Odell, JJ
Øzsu, MT
Papazoglou. MP
Rumbaugh, J
Sheth, A
Shlaer, S
Sølvberg, A
Sowa, JF
Stonebraker, M
Thalheim, B
Ullman, J
Yourdon, E

Table 3.1: List of authors.

The search process resulted in 41 text books, of which 29 were selected for further reviews. Some books were discarded because they focused entirely on specific database languages or database implementation issues only. Other books were discarded because the contents and the authors were more or less the same.

3.2.2 Analysis.

The books were ordered from BIBSYS and reviewed in the order they arrived. Each book was repeatedly and systematically reviewed according to the eleven general ideas developed in chapter 2 and displayed in table 3.2. Guided by the associated indicator terms, indexes and glossaries were scanned for relevant terms. For each relevant term that appeared in the index, pages were looked up and annotations made for later analysis. In cases where the index or glossary did not contain relevant references, the contents were inspected for relevant chapters instead. In addition to taking notes, the results were also collected in a separate table, as shown in table 3.3. See section 6.3.2 on page 147 for a discussion of the study's credibility.

Senses of classification	**Relevant indicator terms**
Classification as a process of expressing mental concepts.	Classification, Classifying, Categorization, Categorizing, Definition, Defining, Concretization, Concretizing
Classification as a system of concept definitions.	Vocabulary, Terminology, Concept system, System of concepts, Universe of Discourse, Domain, Dictionary, Catalogue, Taxonomy, typology
Classification as an identification process.	Classification, Classifying, Identification, Identifying
Classification as a relation between instances and classes.	Classification, Generalization, Concept, Class, Object, Instance, Occurrence
Related concepts	**Relevant indicator terms**
Definition of concept. Distinction between concept intension and concept extension.	Concept, Idea, Intension and extension Mental Model, Schema, Frame
Definition of class. Distinction between classes as sets of objects and classes as data structures.	Class, Set, Group, Type, Collection Entity type, Object type
Definition of object. Distinction between objects as instances of classes and objects as symbol structures.	Object, instance, occurrence, entity, surrogate, membership condition dynamic and multiple classification
Definition of property/attribute. Distinction between defining and describing properties.	Property, Attribute, Constant, Identifier, primary key, candidate key, surrogate key, oid Function, relation, role, data type
Separation of ideas	**Relevant indicator terms**
Distinction between definition of concepts and descriptions of objects.	Concept, Class, Type, Definition, Description
Classification as a prerequisite to conceptual modeling.	Vocabulary, Terminology Concept system, System of concepts, Universe of Discourse, Domain
Distinction between a concept system and a conceptual model.	Model, Conceptual model, Data model, Domain model, Universe of Discourse

Table 3.2: General ideas of classification and associated indicator terms.

3.3 Findings.

Literature references	Senses of classification				Related concepts				Separation of ideas			
	Cp	**Cs**	**Ci**	**Cr**	**Co**	**Cl**	**Ob**	**Pr**	**Dd**	**Cc**	**M**	**N**
Ambler 1995	n	n	n	n	n	n	y	n	n	n	y	**2**
Atzeni, Ceri, and Paraboschi 1999	n	n	n	n	n	n	n	n	n	n	y	**1**
Boman, et.al. 1997	y	y	n	y	y	y	n	y	n	n	y	**7**
Booch 1991	n	n	y	n	n	n	n	n	n	n	n	**1**
Booch, Rumbaugh, & Jacobson 1999	n	n	n	n	n	n	n	n	n	n	n	**0**
Brodie 1984	y	n	y	y	n	n	n	y	n	n	y	**5**
Bubenko & Lindencrona 1984	y	y	y	y	y	y	y	y	n	n	n	**8**
Coad & Yourdon 1991	n	n	n	n	n	n	n	n	n	n	n	**0**
Connolly & Begg 2002	n	n	n	n	n	n	n	n	n	n	n	**0**
Date 2000	n	n	n	n	n	n	n	n	n	n	n	**0**
Elmasri & Navathe 1997	n	n	y	y	n	n	n	y	n	n	n	**3**
Embley 1997	n	n	n	n	n	n	n	n	n	n	n	**0**
Embley, Kurts, and Woodfield 1992	y	y	y	n	y	y	n	n	n	n	n	**5**
Eriksson & Penker 2000	y	n	n	n	n	n	n	n	n	n	y	**2**
Finkelstein 1990	y	y	n	n	n	y	n	n	n	n	y	**4**
Hoffer, George, and Valacich, 2002	n	n	y	n	n	n	n	n	n	n	y	**2**
Jacobson, Booch, & Rumbaugh, 1998	n	n	n	n	n	n	n	n	n	n	y	**1**
Kent 1978	y	y	y	y	y	y	y	y	y	n	y	**10**
Kroenke 2002	y	n	n	n	n	n	n	n	n	n	n	**1**
Lewis, Berstein, & Kife, 2002.	n	n	n	n	n	n	n	n	n	n	n	**0**
Martin & Odell 1992	y	y	y	y	y	y	y	y	n	n	y	**9**
Page-Jones 2000	n	n	n	n	n	n	n	n	n	n	n	**0**
Partridge 1996	n	n	n	y	n	n	n	n	n	n	n	**1**
Robinson & Berrisford 1994	n	n	n	n	n	n	n	n	n	n	n	**0**
Rumbaugh, Jacobson, & Booch 1999	n	n	n	n	n	n	y	n	n	n	n	**1**
Shlaer & Mellor 1988	y	y	y	n	y	n	n	n	y	y	y	**7**
Sølvberg.& Kung 1998	y	y	y	n	y	n	n	y	n	n	y	**6**
Sowa 1984	y	y	y	n	y	y	n	y	n	n	y	**7**
Ullman & Widom 1997	n	n	y	n	n	n	n	n	n	n	n	**1**
Positive scores	**12**	**9**	**12**	**7**	**8**	**7**	**5**	**8**	**2**	**1**	**13**	**84**

Cp	= Classification as a process of defining concepts (y=yes, n=no)
Cs	= Classification as a system of concept definitions (y=yes, n=no)
Ci	= Classification as a process of identification (y=yes, n=no)
Cr	= Classification as a relation between instances and classes (y=yes, n=no)
Co	= Distinguish between concept intension and concept extension (y=yes, n=no)
Cl	= Distinction between classes of objects and classes as data structures (y=yes, n=no)
Ob	= Distinction between unique object instances and multiple object descriptions (y=yes, n=no)
Pr	= Distinction between defining and describing properties (y=yes, n=no)
Dd	= Distinction between definition of concepts and description of objects (y=yes, n=no)
Cc	= Views classification as a prerequisite to conceptual modelling (y=yes, n=no)
M	= Distinction between a model and its constituent concept definitions (y=yes, n=no)
N	= Number of positive scores.

Table 3.3: Findings and coding information.

3.3.1 Classification – a process, a product, a judgement and a relation.

In the context of conceptual modelling, classification can be understood as a process, a product, a judgement and a relation. The four senses are represented by the first four columns in table 3.3. It is possible to call the four senses by different names, but what really matters is that one is aware of the different senses, and that one is able to apply them correctly.

Number of positive scores assigned to books	Number of books	
	N	Frequency
0	12	0.4
1	6	0.2
2	2	0.1
3	6	0.2
4	3	0.1
N	**29**	**1.0**

Table 3.4: Frequency distribution of positive scores on the four senses of classification.

Of the 29 text books, only 3 books came out with 4 positive scores. Among those, only one distinguished clearly between all the four senses (Bubenko and Lindenkrona 1984). For the remaining books that received from 1 to 4 positive scores, the scores were based on quite liberal interpretations of the texts:

In Kent (1978) for example, the first quote below is taken to support the first sense, while the second quote is a more direct expression of the third and fourth sense:

> "If we really did want to define what a data base modelled, we'd have to start thinking in terms of mental reality rather than physical reality. Most things are in the data base because they "exist" in people's minds, without having any "objective" existence. (Which means we very much have to deal with their existing differently in different people's minds.)" (Kent 1978, p. 18).
>
> "A set is determined by a predicate, whose minimal form involves a relationship to an object: the set of things having relationship X to object Y". (Kent 1978, p. 90).

Martin and Odell (1992) and Martin and Odell (1996) do not mention the first two senses of classification directly. However, appendix C in Martin and Odell (1998) contains a layout of what they call a *type glossary* to support the specification of concept definitions and taxonomic relationships. Hence, one liberal interpretation may be that the type glossary implies an understanding of classification as a process of defining concepts, as well as classification as a system of concept definitions.

Another example is Sølvberg and Kung (1998): in section 14.1, which describes the set-theoretic approach to information modeling, the various senses of classification can be derived from a set of formal definitions of entity and relationship classes:

PERSON = {x | x is a person}
MEN = {y | y ε PERSON and the sex of y is male}
WOMEN = {z | z ε PERSON and the sex of z is female}
BOAT = {w | w is a small vessel for traveling across water}
MARRIAGE = {<u,v> | u ε MEN and v ε WOMEN and u,v are married>

These statements may, with some benevolence, be taken to demonstrate:

1. Classification as a system of concept definitions, where broader concepts, such as PERSON, are defined prior to its subordinate concepts MEN and WOMEN.
2. Classification as the process of defining concepts, in the sense where mental concepts must be expressed by intensional definitions.
3. Classification as a process of identification, in the sense that each concept definition may be considered as a test or condition for class membership.
4. A distinction between a model and its constituent concept definitions, in the sense that concepts such as PERSON, MEN and WOMEN must be defined prior to the modeling of the MARRIAGE relationship.
5. A distinction between the *definition* of concepts and *description* of objects, in the sense that the example contains only definitional properties.

Although all of the above interpretations are readily available from the example, none of them are explicitly stated. However, on page 20, there is a statement concerning concepts and models which reads:

> "It is worthwhile to note that a model of the information system's environment must contain a definition of the concepts that are used for talking about the environment. These concepts are also used for specifying the information system model." (Sølvberg & Kung 1998, p. 20).

The quote clearly expresses the idea of classification as a process of defining the concepts. However the statement is not further elaborated with respect to how the concepts should be defined.

Similarly, in Shlaer and Mellor (1988), the first three senses of classification may be derived from a single paragraph. Note that Shlaer and Mellor uses the term 'object' to denote what is more commonly called a class:

> "An *object description* is a short informative statement, which allows one to tell, with certainty, whether or not a particular real world thing is an instance of the object as conceptualized in the information model. An object description must be provided for each object in the model." (Shlaer and Mellor 1988, p. 19).

Nothing is directly said about classification, definitions, concepts, system of concepts, or identification, yet the ideas can be elicited from the text if one knows what to look for.

In Brodie (1984), one has to consider several different definitions of classification, such as:

> "Classification is a simple form of data abstraction in which an object type is defined as a set of instances. This establishes an instance-of relationship between an object type and its instances in the database." (Brodie 1984, p. 33).

> "Classification is a form of abstraction in which a collection of objects is considered a higher level object class. An object class is a precise characterization of all properties shared by each object in the collection". (Brodie and Ridjanovic 1984, p. 281).

In the first definition, classification is understood as a relationship between instances and their class. This is nice and clear.

In the second definition, sentence two may indicate a process of classification, whereby the membership condition (all properties shared by each object) is precisely specified. Hence, one may read into the definition an understanding of classification both as a process and as a product. However, when reading on, it becomes clear already in the next line that the authors speak of *describing* and *identifying* properties, and not *defining* ones:

> "An object is an instance of an object class if it has the properties defined in the class… For example, an object class *employee* that has properties *employee-name*, *employee-number*, and *salary* may have as an instance the object with property values "John Smith", 402, and $50,000". (Brodie and Ridjanovic 1984, p. 281)

The example makes it clear that *to share properties* does not mean to *have the same values* for a set of defining properties, but to *have the same properties* as those specified for the class. That is, according to the example, to be an employee, is to *have* an employee-name, an employee-number and a salary. This is not a very informative definition of employee. It does not tell what it means to be an employee, only what may be said of employees. In reality, this is only a description.

Besides, for identification purposes, this kind of extensional definition becomes more and more useless as the number of properties grow.

Hence, we may draw the conclusion that classes, if they are to be defined by all their properties, must be based on so called intuitive classification. Alternatively, this kind of class presupposes that definitions already exist in order to determine which objects to represent as instances, and which to reject.

The need for precise definitions has also been recognized by Brodie (1984):

> "Concepts that constitute a particular data model must be precisely defined. Precise definitions aid people in understanding the data model, ensuring the soundness of the data model concepts and their interaction, developing analytical tools, and implementing related languages and techniques. Typically, data models have not been formally defined. Consequently, data models are difficult to understand, apply, compare, and analyse". (Brodie 1984, p. 40).

Although one may question Brodie's and Ridjanovic's definition of classification, their view is not uncommon among writers in the field of conceptual modelling. The essence of the view is based on an idea of sameness: Things are grouped into the same class because they are similar.

> "…classification is fundamentally a problem of finding sameness. When we classify, we seek to group things that have a common structure or exhibit a common behavior." (Booch 1991, p. 133).
>
> "Collections of objects share the same types of attributes, relationships and constraints, and by classifying objects we simplify the process of discovering their properties." (Emasri and Navathe 2000, p. 101).
>
> "In an ideal classification, object sets would consist only of objects that all have the same kinds of properties." (Embley 1998).
>
> "…it is usually pointless to classify people, cars and paper clips in one entity type because they have little in common in a typical enterprise model. More useful is classifying semantically similar entities in one entity type, since such entities are likely to have useful common attributes that describe them." Lewis, Bernstein and Kifer, 2002, p. 91).

All four statements represent variations over the same theme: classification is a matter of grouping similar objects into the same class. However, there is a catch here, as demonstrated by the philosopher Goodman:

For any object, an infinite number of properties are potentially relevant to a similarity judgment. The number of properties that plums and lawn movers have in common could be infinite: both weight less that 1,000 kilograms, (and less than 1001 kilograms, and so on), both are found in our solar system, both cannot hear well, both have a smell, both can be dropped, both take up space, and so on.

This seems to imply that all objects are similar to all others. However, all objects will also have infinite sets of properties that are not in common. A plum weights less than one kilogram, while a lawn mover weights more than one kilogram, and so on. This suggests that all objects are dissimilar to all others. For a thorough discussion of the insufficiency of similarity with respect to classification, see (Hahn and Chater 1997; Murphy and Medin 1985; and Barsalou 1992).

Hence, a reference to similarity alone cannot explain why things end up in the same class. For things to end up in the same class, they must be similar *with respect to something* – call it a predicate, a membership condition, a categorization rule, or a principle of classification.

In other words, definitions that refer to sameness or similarity without considering in what *respect* the similarity judgement is meant to be based, are too general to explain what classification means.

As an example of specifying such a *respect*, Finkelstein (1990) speaks of *entity purpose descriptions*:

> “The first task is to define the purpose of each entity; that is, its reason for existence. We do not define how it is used: that will come later when we examine strategies. Rather, we decide what purpose it serves.” (Finkelstein 1990, p. 291).

Purpose may be interpreted as a predicate or as the reason for why a certain predicate has been selected. In that respect, Finkelstein is close to recognize classification as a process and a collection of definitions.

Another example is Embley, Kurts and Woodfield (1992) who give the following definition of classification:

> “Identification of sets of objects that belong together for some logical reason is called *classification*. … An analyst may group any set of objects into an object class for any reason, but the classification should make good sense.” (Embley, Kurts and Woodfield 1992, p. 24).

Here, *logical reason* seems to serve the same function as Finkelstein’s *purpose description*. However, Embley, Kurts and Woodfield (1992) carry the idea one step further by considering membership conditions as well:

> "An object that satisfies the conditions for membership in an object class is a member of the object class. … Object class membership conditions describe objects. For an object they give the object-class name as the class of the object, the generalizations of which the object is a subclass, the direct relationships, the inherited relationships, and the constraints applicable to the object, including participation constraints, co-occurrence constraints, …". (Embley, Kurts and Woodfield (1992, p. 52).

This description of a membership condition seems to be closely mirroring the terminological understanding of a concept system. It demonstrates that membership conditions may be quite complex expressions, that among other things may involve the concept's position in a semantic network

In the remaining text books, Robinson and Berrisford (1994), Ambler (1995), Ullman and Widom (1997), Jacobson, Booch and Rumbaugh (1999), Atzeni et. al. (1999), Eriksson and Penker (2000), Kroenke (2002), Page-Jones (2000), Connolly and Begg (2002), and Hoffer et. al. (2002), classification is not an issue at all, though some books may have received a single score as a result of some liberal interpretations.

What can be said so far is that general ideas pertaining to classification have been found in 60% of the reviewed text-books. However, only in a single book, that is, in 3% of the reviewed material, are all four senses presented as a consistent and interrelated set of ideas. How can this be? One possible answer may be based on a reflection from Martin and Odell's (1992) book:

> "While we can form concepts for which no objects exist, objects cannot exist in a person's awareness without applicable concepts. In other words, a particular object may exist for some people, because they have the conceptual structure necessary to perceive it. However, the same object may simply not exist for others, because it lies outside their set of concepts". (Martin and Odell 1992, p. 236).

Can the answer be that most writers simply lack the necessary concepts to perceive the four senses of classification as 'four senses of classification'? In order to answer this question, it is necessary to study how the authors understand concepts like concept, class, object, and property.

3.3.2 Constructs – concept, class, object and property

To be able to conceive the four senses of classification as a consistent and coherent whole, one must necessarily be well acquainted with some of the more basic concepts like 'concept', 'class', 'object' and 'property'. The concepts and some relevant interpretations are presented in table 3.5.

Concepts	Context of analysis	Context of design
Concept	Intension - Predicates Extension – Class of objects Terms and definitions	Defining attributes, constants, procedures Names, data structures
Class	Collection of objects Analysis classes	Data structures and data Implementation classes Inheritance
Object	Instances Individual identities Typological identities	Multiple descriptions OID's, Surrogates, keys Defining attributes
Property	Defining properties Describing properties	Defining attributes, constants, procedures Describing attributes Identifying attributes

Table 3.5: Basic concepts and interpretations needed to conceive classification.

The two contexts that are displayed in table 3.5 may be understood as two modes of thinking, rather than two separate processes. The context of analysis reflects ideas that are connected to analytical problems, while the context of design reflects ideas that are associated with design solutions in the form of representations and procedures.

For example, when thinking of concepts, if one is aware of the two aspects of intension and extension, it becomes natural to think of possible predicates and how well they serve to identify the relevant sets of objects. One may even think about a name and a definition to express the intension. At the same time, it is natural to reflect on possible representations, not to say implementations. A predicate may for instance be represented by an attribute and become part of a data structure.

Following this argumentation, one may expect that those who score on the four basic concepts are likely to score on the four senses of classification as well. On the other hand, if the four basic concepts are only superficially treated, then classification is most likely not treated at all.

Accordingly, the text books have been reviewed with an eye to the following distinctions:

1. Distinction between concept intension and concept extension.
2. Distinction between a class of objects and a data structure.
3. Distinction between unique object instances and multiple object descriptions.
4. Distinction between defining and describing properties.

These distinctions are not taken literally in the sense that a definition of ‘concept’ for example, necessarily must refer to ‘intension’ and ‘extension’. It is sufficient if the ideas of intension and extension are referred to, for example by terms like ‘predicate’ and ‘entity set’, or ‘criteria’ and ‘object class’.

Concepts and terms

As recorded in table 3.3, eight text books refer to the notion of intension and extension in one way or another. Of these, four books make explicit reference to ‘intension’ and ‘extension’, while the remaining four books refer to the same ideas, but in less obvious manners. This is illustrated by some selected quotes below.

With a few exceptions, the positive entries in table 3.3 coincide with the positive entries already noticed for classification. In addition, and equally important, all negative entries for concepts coincide with the negative entries that are recorded for classification. This suggests that an understanding of concepts in terms of intension and extension seems to be highly correlated with an understanding of classification, as presented in chapter 2.

Below are some selected quotes to demonstrate how concepts are defined, or how their ideas are implied by the text, without being directly named. The first two quotes are examples of the former, the last two of the latter.

Kent (1978) focuses on two distinct notions of ‘set’ to explain intension and extension:

> “There is the abstract idea of what the type is (e.g., the idea of “employee”), and the current population of people who happen to be employees at the moment. The former is the “intension”, and the latter is the “extension”. The latter tends to change often, (as people get hired and fired), but the former doesn’t.” (Kent 1978, p. 90).

Here *type* may be understood as a mental concept, and *population* as a dynamic collection of objects that shrink and grow as objects may be removed and/or added from/to the population. By this interpretation, Kent distinguishes between a ‘type/population’ notion of sets, and a ‘traditional set’ notion that correspond to the traditional axioms of set theory.

The ‘type/population’ notion of sets allows for its extension to change, while the intension remains the same. A mathematical set, however, has no notion of a changing population. If the population changes, so does the set. The implication of this is that the ‘type/population’ notion can serve as a foundation for classification, while the mathematical notion is less suited for that purpose. However, in spite of this implication, the connection from concept to classification, as implied by the quote, is not in any sense obvious to the reader.

In Martin and Odell (1992), a concept is defined as:

> “an idea or notion we share that applies to certain objects in our awareness.” (Martin and Odell 1992, p. 233).

The authors distinguish between privately held ideas, which they call *conceptions*, and shared ideas, which they call *concepts*. In order for conceptions to be agreed upon and shared by others, they need to be concretized and communicated by means of definitions. These reflections are all included in the way the authors define intension and extension:

> “For example, forming the concept Writer requires a clear definition of what it takes to be a writer. Once this definition is in place, we can then identify objects that are instances of the Writer concept. Because we think in this way, concepts are employed as units of knowledge. Adopting this idea has important ramifications. The concept as a unit of knowledge supports discrete definitions of our recognition tests (the intension), and identification of those objects a definition applies to or not (the extension).” (Martin and Odell 1992, p. 237).

The quote expresses ideas that are very similar to those stated in chapter 2, and demonstrates quite clearly how the first three senses of classification may easily be talked about without using the term classification.

In Sølvberg and Kung (1998) there seems to be no explicit definition of ‘concept’. However, the terms ‘intension’ and ‘extension’ are still being used, but then in a more specialized form, especially suited for use in a discourse revolving around object-oriented design issues.

> “An object class defines the structure, i.e., the attributes and their types and the operations. An object class can be interpreted in two different ways: 1) it defines the intension, constraints on the attribute values that an object of the class can have, and constraints on the invocation of operations; 2) it defines the Herbrand universe, i.e., all possible objects of the class. To distinguish between these two interpretations some authors use object type to refer to the first interpretation and object class to refer to the second.” (Sølvberg and Kung 1998, p. 410).

The way the authors use the terms intension and extension does not make the connection between concepts and classification very obvious. Still, the ideas are there if one first knows what to look for.

The final example is taken from Shlaer and Mellor (1988). The authors are very "thing-oriented" in the sense that in their methodology, classes are identified by focusing on things in the problem at hand, rather than on concepts. Consequently they are neither concerned with concepts nor with intension and extension. In spite of this, the ideas underlying concept, intension and extension can easily be recognized by the following expressions:

> "An object is an abstraction of a set of real-world things …". (Shlaer and Mellor, 1988, p. 14).

Here, 'abstraction' can be interpreted to stand for the idea of a concept, or of a concept's intension, while 'a set of real world things' is a common way to denote a concept's extension.

> "An *object description* is a short, informative statement which allows one to tell, with certainty, whether or not a particular real world thing is an instance of the object as conceptualized in the information model." (Shlaer and Mellor, 1988, p. 19).

An object description is referred to by the authors as the basis for abstraction. The whole formulation can be read as a definition of 'concept', which fully depends on the notion of intension and extension for its meaning.

To summarize, 28% of the text books make use of the notion of intension and extension in their definitions of concepts. The results are highly correlated with the results from the previous reviews of classification. Most probably, the explanation for this correlation is that if one defines concepts with reference to intension and extension, then the three first senses of classification follow almost by logical implication.

Classes, sets and types

If the connection between concepts and classification is not evident in the text books, then the connection may still be reflected in the way classes are defined and elaborated. Classes may be defined with reference to *intension* and *extension*. Or, classes of objects may be distinguished from data structures in the sense that classes are associated with *defining* properties, while the data structures are associated with *describing* properties.

If one or more of these aspects are mentioned in the definition of class (or type), then a y(es) is recorded in table 3.3

Seven text books have received a positive entry. Of these, six define a class as a set of objects, which is determined by 'a predicate' (Kent 1978), 'tests' (Martin and Odell 1992), 'a purpose description' (Finkelstein 1990), 'selected properties', (Bubenko and Lindeckrona 1984), 'essential properties' (Sowa 1984), 'logical reason' (Embley, Kurts and Woodfield 1992), or 'respect' (Boman et. al. 1997). In addition, Sowa (1984) also emphasizes the distinction between defining and describing properties:

> "Type definitions present the narrow notion of a concept, and a schemata presents the broad notion. The Aristotelian and Scholastic distinction between essence and accident makes a similar point: type definitions are obligatory conditions that state only the essential properties, but schemata are optional defaults that state the commonly associated accidental properties." (Sowa 1984, p. 128)
>
> "Whereas a type definition for EMPLOYEE presents the primary defining characteristic, a schema would include the background information that an employee has an employee number, earns a salary, reports to a manager, works in a department, and so forth". (Sowa 1984, p. 304).

In table 3.3, one can see that six of the seven positive entries for the notion of 'class' coincide with positive entries already made for the notion of 'concept'. Of the remaining twenty two text books, sixteen define classes based on the notion of sameness, while four text books distinguish between a class as a set and a class as a data structure, but without mentioning anything about defining and describing properties. The last two text books view classes as a design or implementation construction.

Typical examples of class definitions based on the notion of sameness are:

> "A class is a set of objects that share a common structure and a common behaviour." (Booch, 1991, p. 93).
>
> "The real-world concepts represented by the objects of a class should be similar. …The properties of objects in a class must be the same.". (Ullman and Widom, 1997, p. 27-28).
>
> "An entity type defines a collection (or set) of entities that have the same attributes." (Elmasri and Navathe 2000, p. 49).

If we compare these statements with the definitions of classification given at page 71, it becomes clear that definitions of classes based on sameness suffer from the same insufficiency as definitions of classification based on sameness.

Without a statement of what is to count as the principle for a similarity judgement, there is really no way to conceive of a class, collection or set, because any two objects can be as similar or dissimilar as we want them to be.

In chapter 2, several papers on conceptual modelling, such as Smith and Smith (1977), Codd (1979), and Hammer and McLeod (1981) refer to so-called predicate-defined, and user-defined subclasses. The same ideas can be found in Elmasri and Navathe (2000):

> "In general, we may have several specializations defined on the same entity type (or superclass)… In some specializations we can determine exactly the entities that will become members of each subclass by placing a condition on the value of some attribute of the superclass. Such subclasses are called predicate-defined subclasses…
>
> This condition is a constraint specifying that members of the subclass must satisfy the predicate, … If all subclasses in a specialization have the membership condition on the same attribute of the superclass, the subclass itself is called an attribute-defined specialization, and the attribute is called the defining attribute of the specialization… When we do not have a condition for determining membership in a subclass, the subclass is called user-defined. Membership in such a subclass is determined by the database users when they apply the operation to add an entity to the subclass; hence, membership is specified individually for each entity by the user, not by any condition that may be evaluated automatically". (Elmasri and Navathe 1997, p. 80-81).

Why are predicates introduced only for subclasses and not for regular classes, or top level superclasses? Part of the answer may be that predicates are needed to simplify the implementation of generalization hierarchies, including operations on the structures, such as queries, and multiple inserts, updates and deletes. But this does not explain why some classes get away without a predicate. At least in theory, all classes should be associated with a predicate, since it represents the meaning of a class. However, the authors seem to accept the idea of user-defined subclasses, and in that respect, I can only think of one possible answer, and that is a belief in intuitive classification.

To summarize briefly, 24% of the text books refer to the notion of intension when defining classes, but they make use of different terms, such as 'predicate', 'tests', 'purpose description', 'selected properties', 'essential properties', 'logical reason', or 'respect'. Only a single book distinguishes between defining and describing properties.

Six of the seven positive entries for the notion of 'class', coincide with positive entries already made for the notion of 'concept'. Again, the reason for this correlation may be that a definition of concept by means of intension and extension contains a definition of class as one of its aspects.

Objects, things and entities

Based on the notion of concepts and classes, one may expect objects to be defined as instances of concepts, sets, classes or types, and that the idea of typological identity is somehow included in the definition. In addition, since different purposes may cause a single object to be variously described, one may expect a discussion of defining and describing properties, sets of objects and data structures, and of individual and typological identity. Reviews of the text books show that objects, things, or entities, are commonly defined with respect to contexts and/or to certain object characteristics. With respect to contexts, objects are defined as instances of either conceptual classes or as instances of implementation classes. Instances of implementation classes are *symbolic representations* of instances of conceptual classes. As an example, Boman et.al (1997) distinguish between things (as instances of conceptual classes) and their linguistic representations (as instances of implementation classes):

> "An object is a thing or phenomenon in the real world. ... Objects belong to the object system (a part of reality), whereas the object identifiers belong to the language used in reasoning about the object system". (Boman et.al 1997, p. 49).

Though the individual identity of objects is generally recognized, the typological identity is not an issue at all. Multiple descriptions however, have been recognized to some degree. Five text books, Kent (1978), Bubenko and Lindenkrona (1987), Martin and Odell (1992), Ambler (1995), and Rumbaugh, Jacobson, and Booch (1999) have received a positive score by recognizing that an object may belong to more than one class at any one time. Of these, three have already received positive scores on all previous questions. Kent (1978) for example, distinguishes between entities, entity types, and record types. A single entity may be an instance of several entity types, each of which is represented by an associated record type:

> "If we intend to use a record to represent a real world entity, there is some difficulty in equating record types with entity types. It seems reasonable to view a certain person as a single entity (for whom we might wish to have a single record in an integrated database). But such an entity might be an instance of several entity types, such as employee, dependent, customer, stockholder, etc". (Kent 1978, p. 104).

Similarly, Martin and Odell (1992) emphasize the dynamics and multiplicity of classification from a design and run-time perspective. The fact that a single object may belong to different types, and change membership over time, means that the uniqueness of objects must be recognized across classes.

This perspective brings out the same distinctions as those emphasized by Kent: first there are objects, then there are sets to which the objects may enter and/or leave, and finally there are data structures associated with each set whereby its members are further described.

> "In her lifetime, Jane may be a member of several sets, and may, on many occasions, change her set membership. This means, first, that an object can have multiple concepts that apply to it at any one moment. Second, it means that the collection of concepts that applies to an object can change over time". (Martin and Odell 1992, p. 245).

For the remaining twenty four text books, objects are mostly described with reference to one or more of the following: *identity*, *structure*, *behavior*, and *persistence*. It is worth noticing that typological identity is not mentioned at all. Every object is associated with a unique, system-generated object identifier, commonly referred to as *oid* or *surrogate key*. Regarding structure, objects range from *simple* objects such as literals, to *complex* objects composed of collections and/or tuples. Behavior represents the *operations* that can be applied to objects of a certain type, and the lifetime of an object refers to whether an object is *persistent* or *transient*. Persistent objects are stored in the database and persist after program termination. Transient objects exist in the executing program and disappear once the program terminates. (Elmasri and Navathe 2000). A few example definitions are quoted below.

Kroenke (2002) defines an object with reference to identity and structure.

> "A semantic object is a representation of some identifiable thing in the users' work environment. More formally, a semantic object is a named collection of attributes that sufficiently describes a distinct identity." (Kroenke 2002, p. 80).

Booch (1991) adds behavior:

> "An object has state, behavior, and identity; the structure and behavior of similar objects are defined in their common class; the terms instance and object are interchangeable". (Booch 1991, p. 77).

Sølvberg and Kung (1998) add context and encapsulation:

> "An object models an entity or thing in the application domain. For example, books, employees, etc., in the real world can all be modeled by objects.
>
> An object has a set of attribute values that define a state of the object. For example, the status attribute of a library book may have as its values "available", "checkout", "on reserve", "missing", and "removed". These values may be used to determine the state of a book object at any time.

> An object has a set of operations that the object is capable of performing to change its attribute values, and may cause changes to attribute values of other objects. For example, filling an order in a retail company may cause the following changes: 1) the order changes its state from "new order" to "filled order"; 2) the customer's balance is changed to reflect the additional amount that is charged to the customer; and 3) the inventory level or quantities-on-hand of the merchandise is updated to reflect the amount sold to the customer.
>
> An object encapsulates both its attributes and operations; this means that the attributes and operations of an object are modelled and stored together with the object. In the function-oriented and data-oriented paradigms, the attributes and the operations of an object are modelled and stored separately …
>
> An object has identity that can be used to uniquely identify the object, or distinguish the object from other similar objects. Each object has its own identity so that even if two objects have the same attribute values, they can still be identified by using their identities". (Sølvberg and Kung 1998, p. 409-410).

As can be seen from the quotes, nothing is said to connect objects and classification, though there are many possibilities. To use the last quote as an example, ideas pertaining to classification could have been introduced in each of the five paragraphs:

1. Objects as models suggest that one thing may be modeled by different objects depending on purpose.
2. State values suggest meaningful partitions of a class into disjoint subclasses, in which the state attributes serve as principles of classification.
3. Operations are often described with reference to the creation, destruction and updates of objects. Operations to verify class membership could have been discussed in this context.
4. Encapsulation is first of all a design characteristic that veils the distinction between real world objects and implementation objects. The insistence to view an object as a bundle of data and operations makes it easy to think of a one-to-one relationship between an object and its description.
5. Finally, object identity is one kind of identity, but typological identity could have been mentioned as well.

To summarize, 17% of the text books have received a positive score because they have distinguished between objects, sets of objects, and record structures, or because they have discussed aspects of dynamic and multiple classification. The remaining 83% seem to be so concerned with aspects of design and implementation, that conceptual aspects are overlooked.

Characteristics, properties and attributes

The distinction between defining and describing properties is only recognized by a single writer, Sowa (1984):

> "Type definitions present the narrow notion of a concept, and a schemata presents the broad notion. The Aristotelian and Scholastic distinction between essence and accident makes a similar point: type definitions are obligatory conditions that state only the essential properties, but schemata are optional defaults that state the commonly associated accidental properties." (Sowa 1984)

Of the remaining text books, seven books, Kent (1978), Martin and Odell (1992), Sølvberg and Kung (1998), Bubenko and Lindencrona (1984), Elmasri and Navathe (1997), Brodie (1984), Boman et.al. (1997), define attributes as associations between objects:

> An attribute of an object is an identifiable association between the object and some other object or set of objects. …An Attribute type is a function. (Martin and Odell 1998, p. 275).

> The meaning of the elements of a value set is defined by an association of the value set to an entity or relationship class. The association is called an attribute of the entity or relationship class. (Sølvberg and Kung 1998, p. 483).

This view accords with a realist ontology where properties are seen as universals that are related to objects via an *exemplification* relationship, (Grossmann 1992). Since one and the same universal property can be exemplified by many objects at the same time, nothing is better suited to serve classification than universal properties.

It is interesting to see that the same text books have scored positively on a number of previous questions.

In eight other text books, Shlaer and Mellor (1988), Coad and Yourdon (1991), Ullman and Widom (1997), Finkelstein (1990), Kroenke (2002), Connolly and Begg (2002), Lewis, Bernstein and Kifer (2002), Ambler (1995), attributes are defined by saying that they describe objects:

> "The particular properties of entity types are called attributes. For example a Staff entity type may be described by the staffNo, name, position, and salary attributes. The attributes hold values that describe each entity occurrence and represent the main part of the data stored in the database". (Connolly and Begg 2002, p. 338).

> "Entities have attributes or, as they are sometimes called, properties that describe the entity's characteristics". (Kroenke 2002, p. 52).

It would be too much to claim that this view represents a nominalist ontology, but to view properties solely as descriptive or identifying properties is just as meaningless to classification as nominalism is. This is also reflected in table 3.3. Text books in which attributes are defined as describing properties only, have less positive scores on the previous questions than have those where attributes are defined as associations.

The remaining 14 text books, define properties in ways that are totally irrelevant for this study, for instance by defining an attribute as “a column in a table”, or as “something of interest to the organization”.

To summarize, only a single text book, that is, 3% have noticed the distinction between defining and describing properties. However, positive scores were also allotted to those writers who define properties as associations between objects, simply because this view explains the basic principles of classification so well.

3.3.3 Classification versus conceptual modeling

Classification can be distinguished from conceptual modelling, as can vocabularies from conceptual models: the classification process produces vocabularies which names and defines the concepts with which conceptual models are built. In this sense classification has, according to Dunnell (1994), primacy over structures, structuring, models, and model-building, in the sense that one must first select and disambiguate the pieces before one can build something out of them. Another way of putting this is to say that a vocabulary determines the meaning of a set of terms, but not how the terms are going to be used in the conceptual model. For instance, a term like ‘Student assistant’ may be defined as a subordinate term to the term ‘Person’. However, when it comes to the conceptual model, the term may be used in several ways: it may be used to denote a subordinate entity type, or a role, or an attribute of the Person entity type, or even a value in a value set associated with the Person entity type.

Further, classification, in the sense of identification, has primacy over description. One must first identify an object as an instance of a class before it can be further inspected and described.

Alternatively classification and conceptual modeling may be viewed as two sides of the same coin: on the one hand, classification may suggest some structures to be directly copied by the conceptual model while on the other hand, the model constructs of the conceptual model may influence the selection and definition of terms.

Possibly, the vocabulary and the conceptual model are best developed in parallel, but still, there is a distinction that may be reflected in the text books. Therefore, the purpose of this final review is to find out to what extent the authors are aware of any distinctions between definition of concepts and descriptions of objects, or between classification and conceptual modelling, or between a terminological concept system and a conceptual model.

Close to 50% of the reviewed text books have received a positive score on the final question regarding the distinction between a model and its constituent concept definitions. Most of the scores were given because of the authors' concern for glossaries. For example, in Eriksson and Penker (2000) an interesting section presents of a pattern for term definitions, which can be used to document and analyze terminology for large enterprises. The main idea behind the pattern is that critical concepts within the business must be unambiguously defined by means of a term, its usage, and its meaning.

In Jacobson, Booch, & Rumbough, (1999), domain modelling is suggested as a means to express the system context as part of the requirements capture:

> "The purpose of domain modeling is to understand and describe the most important classes within the context of the domain… The glossary and domain model help users, customers, developers, and other stakeholders use a common vocabulary. Common terminology is necessary to share knowledge with others. Where confusion abounds, engineering is difficult, if not impossible". (Jacobson, Booch, and Rumbough, 1999, p. 121)

The importance of developing a glossary of terms is clearly stated, but otherwise, there is no follow up on how the glossary should be created. Any distinctions between concepts and objects or between intensions and extensions are not mentioned, and notions of concepts and classification is neither formally defined nor reflected in the text.

Similarly, in appendix C in Martin & Odell (1998) a layout of a type glossary is specified to support the specification of concept definitions. Hence, a positive score has been given to the last question.

However, it is not clear which of the glossary and the data model should be developed first, or whether they should be developed in parallel.

In contrast, Shlaer & Mellor (1988) are very clear on this issue:

> "The fact that these separate vocabularies and, more significantly, their implied *separate conceptual frameworks* exist in an organization should be taken seriously: One has to assume that the subject matter is sufficiently complex that a single vocabulary could not arise through normal informal processes. As a result, real intellectual effort is required for investigating and resolving possible differences. Until this is done, any attempt at stating system requirements is bound to be troubled, since no one can be certain exactly what vocabulary has been used in the requirement statement." (Shlaer and Mellor 1988, p. 2).

According to the quote above the authors seem to suggest that classification is a prerequisite to conceptual modelling. In addition, they also speak of object descriptions that distinguish between defining and describing properties:

> "A short informative statement, which allows one to tell, with certainty, whether or not a particular real world thing is an instance of the object as conceptualized in the information model". (Shlaer and Mellor 1988, p. 19).

Boman et. al. are also very clear, but they seem to take the opposite view that the glossary and the conceptual model are developed in parallel and constitute a single product:

> "Conceptual modelling can make it easier for the actors of an organization to arrive at consensus on a common world view, to use the same language, and to agree on rules that should prevail in the organization. The help conceptual modelling provides is a clear definition of concepts, their properties and relationships, as well as clear definitions of the dynamics of such systems. Clear definitions help to detect disagreement and to arrive at consensus." (Boman et. al. 1997, p.15).

To summarize, close to 50% of the text books recognize a distinction between classification and conceptual modelling, mostly by emphasizing the importance of developing a glossary of terms. Although glossaries or dictionaries are commonly seen as something distinct from the conceptual model, the classification process is generally recognized as something integral to the modelling process. Only a single book, Shlaer and Mellor (1988), views classification as a separate process to be conducted prior to the modelling process.

3.4 Overall summary and conclusions.

This study tests the assumptions stated in section 3.1, that the notion of classification is not sufficiently attended to by the data modeling community, and that a lack of attention causes related notions, such as concept, class, object and property to be unclear, ambiguous and inconsistent in their definitions.

To demonstrate that classification is not *sufficiently attended to* depends on what *sufficiently attended to* means. However, because of the interpretational nature of this inquiry, crisply defined criteria may lessen the validity of the results by excluding relevant expressions and examples from being considered. Accordingly the term can be allotted a more flexible role.

1. If *sufficiently attended* to is taken to mean that all four senses of classification must be explicitly defined, then there is not a single text book in the sample that satisfies the requirement. That is, classification is not attended to at all.
2. If *sufficiently attended* to is taken to mean that the four senses of classification should be consistently presented, but not necessarily explicitly defined, then a total of three text book, that is, 10% of the reviewed material satisfies the requirement.
3. If *sufficiently attended to* is taken to mean that, based on liberal interpretations of the texts, one should be able to recognize at least one of the senses of classification, then classification is sufficiently attended to by 17 text books, or 60% of the sample.

The third alternative can be rejected on the grounds that to know a single sense is insufficient all the time classification is known to have, not only several senses, but *several and related* senses.

The second alternative can also be rejected on the grounds that it depends on interpretations of texts. If it is necessary to engage in active text interpretation in order to learn about classification, then classification is not sufficiently attended to.

This leaves us with the first alternative, that the four senses of classification must be explicitly defined. As the only, viable alternative, this leads to the conclusion that the first assumption is supported: The notion of classification is not sufficiently attended to by the data modelling community.

The second assumption is also supported in the sense that the data indicate a positive correlation between classification and the definition of related concepts such as concept, class, object and property. In cases where all four senses of classification have been recognized, the definitions of concept, class, object and property seem to be more consistent as well.

Lastly, close to 50% of the surveyed authors, through their text books, recognize a distinction between a conceptual model and its constituent concept definitions. There seems to be a common understanding that during conceptual modeling concepts need to be negotiated, agreed upon, formalized and unambiguously defined. The result is variously known as a dictionary, data catalogue, glossary, vocabulary, terminology, or even ontology. However, except for the common knowledge that a vocabulary may reduce confusion when discussing systems requirements, or interpreting conceptual models, several questions are generally left unanswered:

1. What purposes are vocabularies meant to serve?
2. What does it mean to formalize and disambiguate terms?
3. Why is it so important to formalize and disambiguate terms?
4. Practically, how are formalization and disambiguation done?
5. What are the general principles for vocabulary construction?
6. Are vocabularies and conceptual models separate products or a single product?
7. Do the terms ‘conceptual model’ and ‘vocabulary’ mean the same thing?
8. Are vocabularies developed prior to the conceptual model, after the conceptual model, or in parallel with the conceptual model?

These are questions that need to be explained in order for designers to apply and benefit from classification during conceptual modelling. The questions are further pursued in the next chapter.

4.0 Including classification concepts in conceptual modeling – methodological aspects

4.1 Introduction.

This chapter describes a methodology for including classification concepts in conceptual modeling. Before presenting the methodology, section 4.2 discusses the nature of conceptual modeling with respect to its underlying ontological, epistemological, and methodological assumptions. In section 4.3, some theoretical aspects are briefly repeated and simplified, primarily to serve as an informal introduction to the methodology in section 4.4. Further details about the theoretical aspects in section 4.3 are covered by chapter 2.

In chapter 2, two major views on classification were identified. One view, called *intuitive classification*, holds that classes exist in the world and that they can be discovered by following simple, heuristic guidelines, such as looking for people, things, roles, interactions, and places. Based on the findings in chapter 3, it seems as if this view is currently the most dominant one in conceptual modeling. The other view is more complex, and involves an interplay among cognitive, linguistic, and ontological elements. Here, classification is understood as a social process between users and designers, where mental concepts are concretized and reconciled into a common vocabulary.

To find out which of the two views are best suited for conceptual modeling, section 4.2 takes a closer look at the nature of conceptual modeling.

4.2. Positivist versus constructivist perspectives on conceptual modeling.

Most text books in data modelling and database theory contain contradictory statements about the nature of conceptual modelling. On the one hand, data models and data bases are described in purely *positivist* terms, while the associated methodology and design heuristics seem to be firmly grounded in *constructivist* ideas.

To exemplify, consider the following statements about modelling formalisms, conceptual models and databases:

> The basic object that the ER-model represents is an entity, which is a «thing» in the real world with an independent existence (Elmasri & Navathe, 2000, p. 45).

> The entity-based approaches to data modeling tend to follow in the footsteps of the objectivist tradition. Under this interpretation, a DM is like a mirror or picture of reality. Reality is given 'out there', and made up of discrete chunks which are called entities. Entities have properties or attributes. Both entities and their properties have an objective existence. (Klein and Hirschheim 1987, p. 10).
>
> Data models enable us to capture, partially, the meaning of data as related to the complete meaning of the world. (Tsichritzis & Lochovsky 1982, p. 6).
>
> The most important basis for developing the logical design of a database is the *real world* of the using environment… The database reflects an image of *the real world*. (Gordon C. Everest 1986, p. 199)
>
> We begin by describing *the world* in terms of entity types that are related to one another in various ways. (Edmond 1992, p. 241)
>
> A database represents some aspects of *the real world*, sometimes called the miniworld or the Universe of Discourse. (Elmasri & Navathe 2000, p. 4)

These statements suggest that categories carve nature at its joints, and that categories are there to be *discovered* by the analyst/designer. According to Guba (1990), this amounts to a *positivist* position which assumes a *realist* ontology. That is, the belief that there exists a reality, driven by immutable natural laws, and that the business of science is to discover the nature of reality. Being committed to a realist ontology, one is constrained to practice an *objectivist* epistemology. The investigator and the investigated object are assumed to be independent entities, and the investigator to be capable of studying the object without influencing it or being influenced by it.

However, in the same books, these positive statements are usually mixed with statements of a radically different flavour, emphasizing the user's *perspectives*, the designer's responsibility of *interaction* and *communication* with the users, and a recommended *hermeneutic/dialectic approach* to the subject matter:

> The central objective of the logical database design process is to model the collective user perceptions of the real world (Gordon C. Everest 1986, p. 199).
>
> Large software systems generally require the integration of diverse specializations: Financial managers, auditors, engineers, operations experts, and the like must be drawn into the requirements and analysis process. In attempting to do this one typically finds areas of partially-overlapping expert knowledge, each with separate and sometimes conflictive vocabularies. … The fact that these separate vocabularies and, more significantly, their implied separate conceptual frameworks exist in an organization should be taken seriously (Shlaer & Mellor 1988, p. 2).
>
> A database typically has many users, each of whom may require a different perspective or view of the database (Elmasri & Navathe 2000, p. 10).

> It is the responsibility of database designers to communicate with all prospective database users, in order to understand their requirements, and come up with a design that meets these requirements. … Database designers typically interact with each potential group of users and come up with a view of the database that meets the data and processing requirements of this group. These views are then analyzed and integrated with the views of other user groups. The final database design must be capable of supporting the requirements of all user groups (Elmasri & Navathe 2000, p. 12).
>
> Forms reveal how somebody thinks about the problem. Read both blank and completed forms, and investigate discrepancies/irregularities (Schlaer & Mellor 1988, p. 86).
>
> Tabulations of data can also be an interesting source. As with forms, you may find poorly factored attributes, cases of misattribution, and similar flaws. Try to unearth the underlying assumptions (Schlaer & Mellor 1988, p. 87).
>
> Dialog is an ancient technique, still unsurpassed, in which thinking men seeks to state valid and universally applicable definitions ('universal truths'). An effective dialogue typically swings back and forth between high levels of abstraction and intensive examination of mundane examples…. The participants in the dialog usually include the experts or specialists in the field being explored as well as modeling specialists. Both roles are required. Frequently, as the dialogue continues, we see specialists interchanging roles. This is an indication that things are going well: The modelers are gaining enough understanding to really talk with the experts (Schlaer & Mellor 1988, p. 87).

Contrary to the first statements, which suggest a positivist position, these latter ones seem to suggest a constructivist position, which, according to Guba (1990), Guba & Lincoln (1998), holds a *relativist* ontology, a *subjectivist* epistemology and a *hermeneutic and dialectic* methodology.

Realities are apprehensible in the form of multiple mental constructions, socially and experientially based, local and specific in nature, dependent for their form or content on the persons or groups who hold them. This forces the constructivist to choose a subjectivist epistemological position, because it is the only way to elicit the constructions held by individuals. If realities exist only in the respondents' minds, interaction seems to be the only way to access them.

Methodologically then, the constructivist proceeds in ways that aim to identify the variety of constructions that exist and bring them into as much consensus as possible. This process has two aspects: hermeneutics and dialectics. The hermeneutic aspect consists in depicting the individual constructions as accurately as possible, while the dialectic aspect consists of comparing and contrasting these existing individual constructions so that each respondent must confront the constructions of others and come to terms with them.

Given these two views of positivism and constructivism, which of the two are best suited as a paradigm for conceptual modelling? According to Guba & Lincoln (1998), it all depends on how the discipline responds to three simple questions:

1. *The ontological question*. What is the form and nature of reality and, what can be known about it?
2. *The epistemological question*. What is the nature of the relationship between the knower and what can be known?
3. *The methodological question*. How can the inquirer go about finding out whatever he or she believes can be known?

4.2.1 What is the form and nature of the reality which conceptual models are supposed to model?

According to Borgida, Mylopoulos and Wong (1986), Mylopoulos (1998), and Allen (1997), the notion of conceptual models is obtained from the cognitive sciences and their concerns with mental models. Since mental models are not directly observable, conceptual models have been introduced to symbolically represent mental models. Hence, in Borgida, Mylopoulos and Wong (1986), *conceptual modelling* is defined as the specification of models that are closer to the human's conceptions of reality than to the machine's representation, and a *conceptual model*, is defined as a number of symbol structures and symbol structure manipulators, which are supposed to correspond to the *conceptualizations* of the world by human observers. In contrast, a *data model* is defined as a specification of the rules according to which data are structured and what associated operations are permitted on them.

Generally speaking, *conceptualizations,* or *conceptions* (Martin and Odell 1998), refers to some hypothesized mental constructs that have their roots in epistemological methods for organizing knowledge, such as classification and instantiation, aggregation and decomposition, generalization and specialization, (Borgida, Mylopoulos and Wong 1986; Coad and Yourdon 1991; and Boman et. al. 1997).

According to Schwandt (1998), conceptualizations are extensively shared, and some of those shared are disciplined constructions, that is, collective and systematic attempts to come to common agreements about a state of affairs, as for example science. This means that a conceptual model is supposed to model, not only the users' and designers' private, and subjectively held conceptualizations, but shared conceptualizations as well, which may be elicited from publicly available materials, such as forms, data sheets, old computerized systems, vocabularies, business models and scientific theories.

A distinction needs to be made between a conceptual model as a *modelling formalism* to represent how people conceive of the world, (Kim and March 1995), and a conceptual model as a *description* that results from applying a modelling formalism.

As a modeling formalism, the model consists of some basic symbol structures that are assumed to correspond to the mental constructs that humans employ to conceive of, and manipulate their perceptions. In this respect, a model is not a passive medium for describing our conceptions. It is a theory about human cognition that shapes our conceptions, and limits our perceptions, (Kent 1978).

As a description, the model is a symbolic representation of the conceptualizations that some people may have of some portion of the world. It is a model of the users' mental models so to say, (Kent 1978), (Kroenke 2002). Since the model constructs are assumed to correspond with human conceptualizations, and since such conceptualization is associated with a few, simple principles of knowledge organization, models in this sense are usually considered to be highly structured, rigid, unambiguous, and simplistic, but not in any way perfect:

> Since the world is a continuum and concepts are discrete, a network of concepts can never be a perfect model of the world. At best, it can only be a workable approximation. (Sowa 1984, p. 345).

In this latter sense, conceptual models are used to facilitate communication between users and designers, to gain insights into the application domain, to analyze and validate information- and transaction requirements, to reason about possible designs, and to document the system (Sølvberg and Kung, 1998).

The documentation which is usually associated with a conceptual model consists of a structured description called a conceptual schema, a diagrammatic presentation of concepts and relationships, and a data catalogue of some kind, in which the concepts and relationships are further described. With respect to classification, it is natural to consider the vocabulary as part of the documentation.

4.2.2 What is the relationship between the designer and the users' mental models?

By assuming that conceptual models are supposed to model mental constructions, a dualist and objectivist epistemology is clearly out of the question. Because of the variable and personal nature of the user's mental realities, individual constructions can only be elicited and refined through interaction between the designers and users, and not from a distant and non-interactive posture taken by the designers alone.

This view is very close to the subjectivist epistemology of constructivism. Here, the investigator and the object of investigation are assumed to be interactively linked, so that the findings are literally created as the investigation proceeds.

> Inquirer and inquired into are fused into a single monistic entity. Findings are literally the creation of the process of interaction between the two (Guba & Lincoln 1998, p. 27).

An important side effect from this posture is that it challenges the traditional distinction between ontology and epistemology; what can be known is inextricably intertwined with the interaction between a particular investigator and a particular object or group.

This may also explain why different designers tend to come up with different models of the same domain, or why users are inclined to validate and accept different models.

4.2.3 How is the conceptual modelling process conducted?

As suggested by the quotations on page 88, conceptual modelling is a combined hermeneutic and dialectic process, which is both shaped and constrained by the users' and designers' conceptualizations, as well as by the modelling formalism being used. This can be illustrated as in figure 4.1:

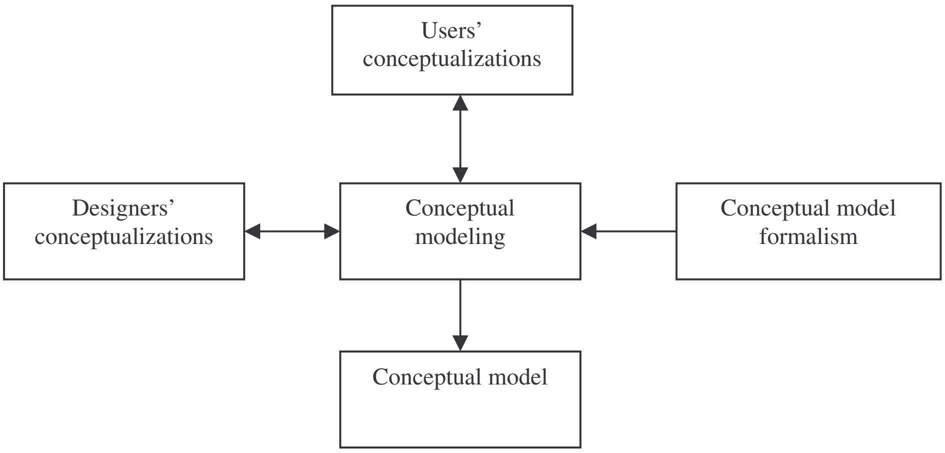

Users' conceptualizations
Designers' conceptualizations
Conceptual modeling
Conceptual model formalism
Conceptual model

Figure 4.1: The conceptual data modelling process.

Users may have different backgrounds and interests, their conceptualizations may be partially private and partially shared, and their vocabularies may be separate and conflicting. Designers may be more or less experienced with respect to the application domain, the conceptual model formalism, and the data model, according to which the system eventually will be implemented.

Lastly, the formalism itself may influence the process with its rigidity and limited collection of constructs, constraints and operations.

During the modelling process the various conceptualizations are interpreted using hermeneutic techniques, and the results compared and contrasted through a dialectic interchange. The final aim of this hermeneutic/dialectic process is to merge conflictive mental models, or conceptual frameworks, into a shared representation, which itself is a conceptual framework, only more informed and sophisticated than the specific predecessor constructions. This approach closely parallels the hermeneutic/dialectic methodology of constructivism:

> Individual constructions are elicited and refined hermeneutically, and compared and contrasted dialectically, with the aim of generating one (or a few) constructions on which there is substantial consensus (Guba & Lincoln 1998, p. 27).

It is important to notice that by this hermeneutic/dialectic approach, conceptual modelling contributes to the construction of new, shared conceptualizations, such as vocabularies, forms, even the whole conceptual model in the end.

4.2.4 Conclusion

Based on the answers given to the three previous questions, conceptual modelling seems to be guided by a constructivist paradigm. If it really is such that conceptual models are supposed to model the users' mental models, and the goal is to arrive at a single, reconciled, symbolic representation, then, an argument from chapter 2, where categorization and classification were viewed as two different mechanisms for establishing order, is well suited to be repeated.

In chapter 2, categorization was used to denote *a cognitive process* of constructing order out of individual, instable, subjective, day to day sense impressions, while classification was used to denote *a social process* of structuring a specific knowledge domain, in order to ensure consistency and stability of meaning between individuals. In order for mental concepts to be talked about, negotiated, and shared, their vagueness, instability and subjectivity were used as a *justification* for classification. Similarly, a constructivist nature of conceptual modelling appears to be a justification for classification as a social process.

Consequently the constructivist nature of conceptual modelling, as developed in this section, favours the view developed in chapter 2, that classification is *a social process between users and designers, where concepts are concretized and reconciled into a common vocabulary.* From this perspective, the notion of intuitive classification is unsuited for conceptual modelling and may be rejected.

4.3 Conceptual modelling – theoretical aspects.

Conceptual modelling can be understood as a social process which aims at a *reconciled* and *formal* representation of an application domain. Depending on the target application, the inputs to the process can vary from highly private conceptualizations of new and innovative ideas, to highly shared and well structured conceptualizations of manual systems, or of computerized systems that are to be rebuilt and modernized.

To be reconciled, means that the users must understand and agree on a single, conceptual model. The model may consist of a controlled vocabulary, one or more diagrams, and one or more schema specifications. The vocabulary may already exist, and part of the user requirements may be to use exactly that vocabulary. Or, the vocabulary may be developed during the conceptual modelling process. Vocabulary may be completed in advance, or developed in parallel with other deliverables of the modelling process. The important point to make is that in the end, the vocabulary must be complete with respect to the terms used in the user requirements, conceptual diagrams and schemas.

To be formal, means that the representations must be *unambiguous*, *precise*, and *formalizable*. To be unambiguous, the vocabulary must be *complete* with respect to the coverage of terms used in the information and transaction requirements. In addition, the term definitions must be *logically consistent* with respect to all other term definitions. Lastly, a term may only be assigned to a single concept, and given concept may only be assigned to a single term. To terminologists, this condition is called *monosemy*, and the set of terms is either called a terminology or a controlled vocabulary.

To be precise, means that the definition of each term must be operationalized, so that the term is applied the same way by different users, as well as by the same user at different occasions. This may require that the definition is complemented with a procedure, illustration or other kind of secondary information to guide the users in their classificatory judgements. The level of precision may vary from term to term, and is determined by the consequences of applying the term incorrectly.

Lastly, to be formalizable, means that it must be possible to map a conceptual schema into a corresponding, logical data model schema, by means of a set of algorithms.

The complete modelling process, from private and shared conceptualizations to the final database schema is illustrated in figure 4.2.

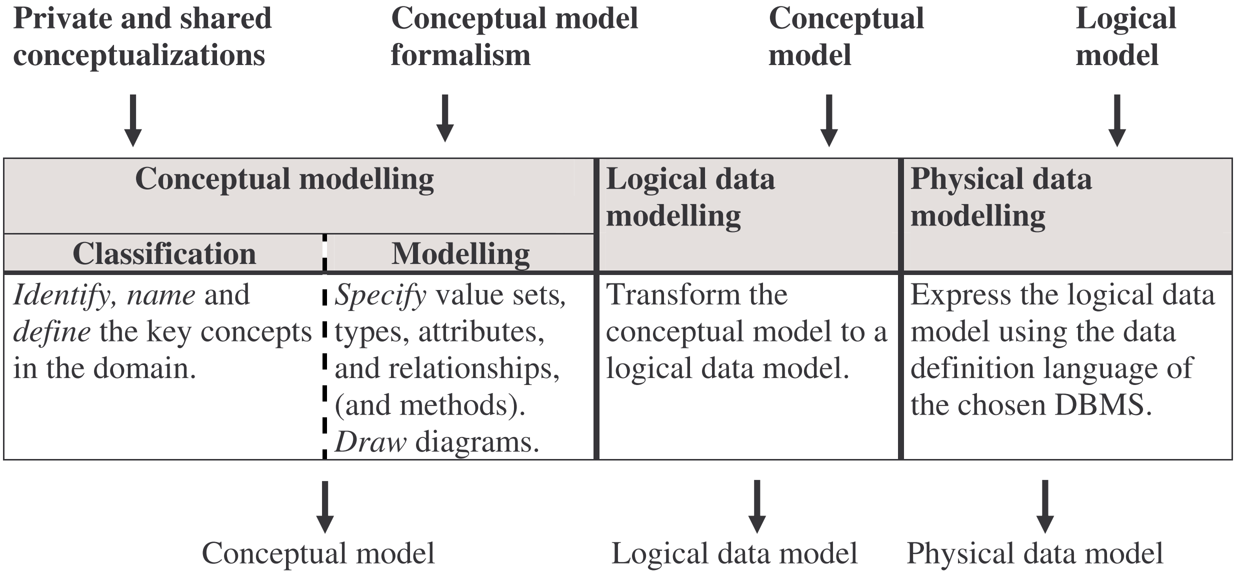

Private and shared conceptualizations
Conceptual model formalism
Conceptual model
Logical model
Conceptual modelling
Classification
Modelling
Logical data modelling
Physical data modelling
Identify, name and define the key concepts in the domain.
Specify value sets, types, attributes, and relationships, (and methods). Draw diagrams.
Transform the conceptual model to a logical data model.
Express the logical data model using the data definition language of the chosen DBMS.
Conceptual model
Logical data model
Physical data model

Figure 4.2: The conceptual data modelling process.

As shown in figure 4.2, conceptual modelling takes private and shared conceptualizations, or user requirements, together with a conceptual modelling formalism as input, and produces a conceptual model as output. The conceptual model, in turn, is input to a process called logical data modelling, where the conceptual model is transformed into a logical data model. The logical data model is then used as input to the physical data modelling process and translated into a physical data model, or database schema.

In figure 4.2, conceptual modelling is divided into two sub-processes, called *classification* and *modelling*. This should not be taken to imply that classification needs to be completed before the actual modelling takes place. Classification and modelling are best considered as two sides of the same coin. Classification is aimed at the *identification, naming* and *definition* of the key concepts in the domain. Modelling is directed towards the *specification* of types, their attributes, relationships, methods, and the *creation* of diagrams. Classification and modelling may be worked upon in parallel or in any sequence, but the final result is not complete before both processes are completed.

As a preparation to section 4.4, the remainder of section 4.3 will mainly focus on the theoretical aspects of classification, briefly repeated from chapter 2. A few basic concepts will be presented, classification will be contrasted with modelling, and the benefits that may follow from classification will be considered.

The practical aspects of classification will be presented in section 4.4.

4.3.1 Basic concepts associated with classification.

Conception

A conception is a mental idea or notion that is associated with an *intension* and an *extension*. The intension states a condition that makes it possible to conceive of a collection of objects as being of the same kind. The extension, is the collection of objects that satisfy the condition. For example, I may want to collect coins when I visit foreign countries. So next time I go to the United States, I may conceive of US coins as the coins I receive when I go shopping. Though my private and subjective conception of US coins is vague and informal, it is sufficiently precise to serve my purpose.

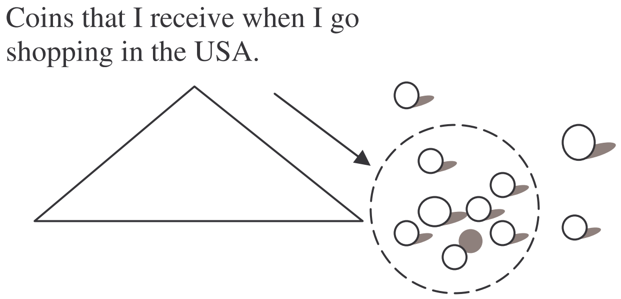


Figure 4.3: A conception, its intension and extension.

Concept

A concept is a conceptualization that is made public and thereby available to others. A given concept is represented by a *terminological entry*. As concepts become available to others, private conceptions may be adjusted to conform to the shared ones.
For example, if I become a more serious collector of coins, I may learn a shared and more precise concept of US coins as illustrated in figure 4.4.

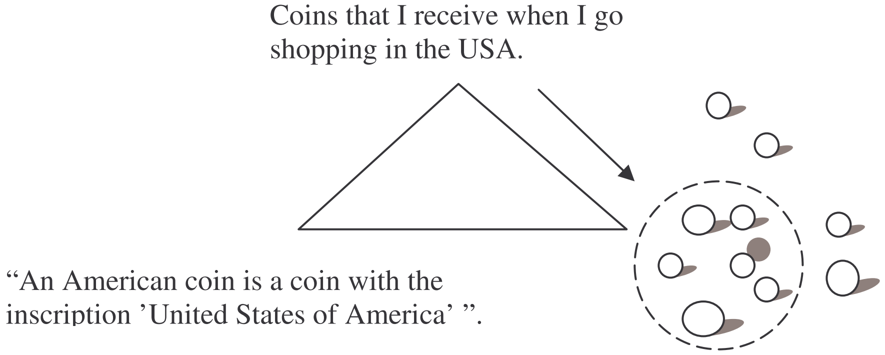

Coins that I receive when I go
shopping in the USA.
“An American coin is a coin with the
inscription ’United States of America’ ”.

Figure 4.4: A concept and a conception, its intension and extension.

As I learn to apply the new concept, it will probably replace my initial informal conception as shown in figure 4.5.

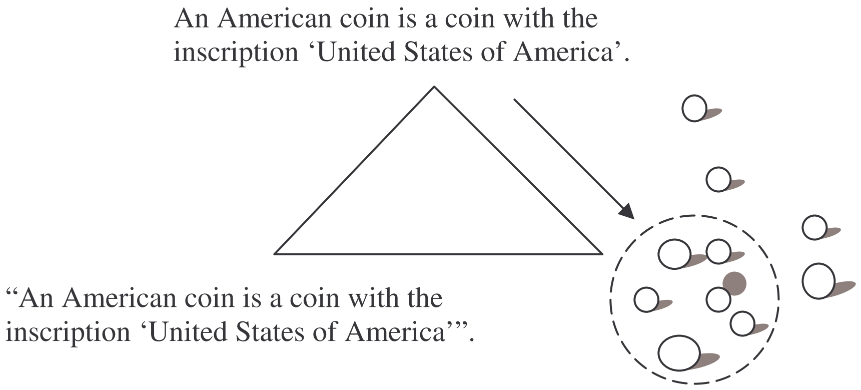

An American coin is a coin with the inscription 'United States of America'.
"An American coin is a coin with the inscription 'United States of America'".

Figure 4.5: A modified conception, based on a concept, its intension and extension.

Terminological entry

Concepts are represented by *terminological entries*. A terminological entry is a statement that explains what a concept means in a certain context. The statement consists of a *subject*, a *copula* and a *predicate*. The subject is a term that works as a shorthand notation for the concept. The copula is understood to be the verb "is", and the predicate constitute the definition, which expresses the concept's intension:

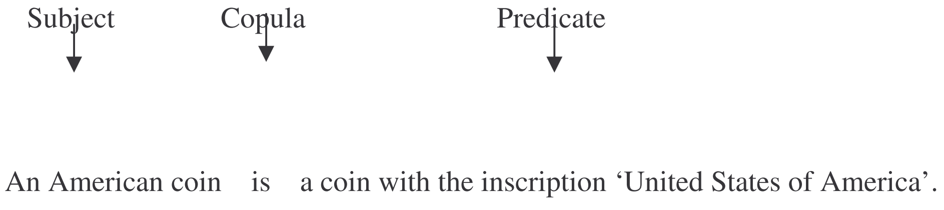


Figure 4.6: The form of a terminological entry.

Term

A term is a designator consisting of one or more words. A term represents a shorthand notation for a concept, and should ideally be a synthesis of the predicate.

Definition

A definition shall express the intension of a concept. Ideally a definition shall indicate a superordinate concept to situate the concept in its proper context, followed by the characteristics that distinguish the concept from other concepts. As demonstrated by the example above, the definition indicate that the concept is to be understood within the context of coins, and that the characteristic that distinguishes US coins from other coins is that they all have the inscription '*United States of America'*.

If intensional definitions become too complex, extensional definitions may be used instead. An extensional definition lists the subordinate concepts in only one dimension as shown below:

Intensional definition:

An American coin is a coin with the inscription "United States of America".

Extensional definition:

An American coin is either a Penny, Nickel, Dime, Quarter, half dollar, or a silver dollar.

Figure 4.7: Examples of intensional and extensional definitions.

Type

A type is the model equivalent of a concept. While a concept states the condition that objects must meet to be identified as instances of the concept, a type is a specification of a logical data structure to describe the instances so identified. Type specifications may be extended to include membership conditions and operations (methods). For further details, see section 4.3.2 on pages 100-101.

Classification

In the context of conceptual modelling, we may distinguish between five senses of classification:

1. *To construct a conception.* Classification is used to denote the cognitive processes that lead to the creation of a conception. For example, to collect US coins, I need to learn how to distinguish US coins from other coins. I may create my own rule, or I may learn a rule from a collector's handbook for instance.

2. *To define a concept*. Classification is also used to denote the process of defining concepts, that is, creating terminological entries.

3. *The system of concept definitions that result from the definition process*. Because most definitions contain a reference to a superordinate concept, the resulting system of concepts will normally be hierarchically organized. The hierarchical structure among concepts may be reflected with numeric labels or with indentation as demonstrated in figure 4.8 and 4.9:

0001 A **means of payment** is a ...

00011 A **coin** is a means of payment that …

000111 An **American coin** is a coin that has the inscription 'United States of …'

000112 A **Greek coin** is a coin that has the inscription 'ΔΡΑΧΜΕΣ'

00012 A **note** is a means of payment that …

000121 An **American note** is a note that …

000122 A **Greek note** is a note that …

Figure 4.8: Classification as a system of labelled concepts.

- A **means of payment** is a ...
    - A **coin** is a means of payment that …
        - **An American coin** is a coin that has the inscription 'United States of …'
        - **A Greek coin** is a coin that has the inscription 'ΔΡΑΧΜΕΣ'
    - A **note** is a means of payment that …
        - **An American note** is a note that …
        - **A Greek note** is a note that …

Figure 4.9: Classification as a system of indented concepts.

4. *The judgement that must be exercised to determine whether an object is an instance of the concept.* This is sometimes called identification instead of classification. For example, is this an American coin?

The answer depends on the definition of American coins. If it is a coin, and, if it has the inscription *"United States of America"*, then the answer is yes, Otherwise, not.

5. *The relationship between instances and their respective classes*. Classification is used to denote the classification between instances and their classes, whereas generalization is used to denote relationships between classes. This sense is illustrated in figure 4.10 below.

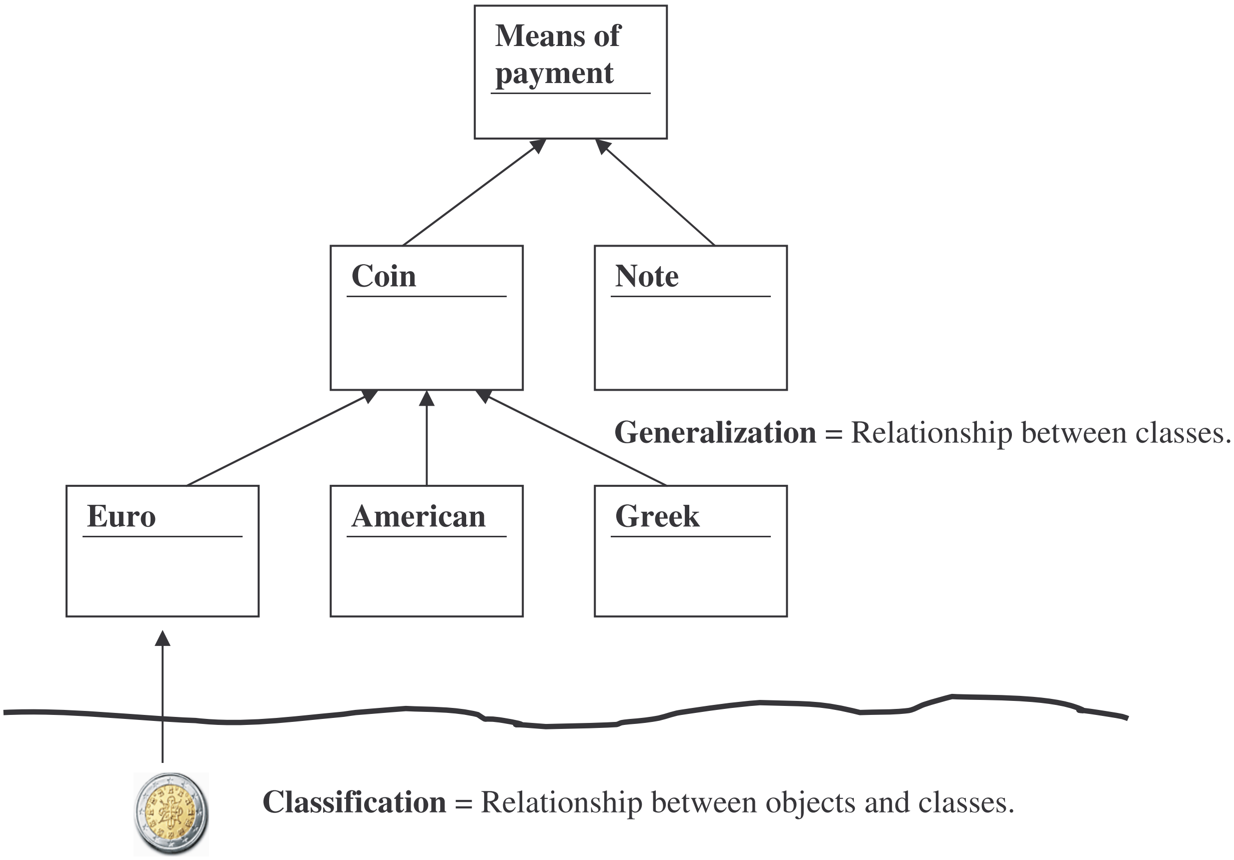

Means of payment
Coin
Note
Generalization = Relationship between classes.
Euro
American
Greek
Classification = Relationship between objects and classes.

Figure 4.10: Classification versus generalization.

4.3.2 Classification versus modelling

Classification is concerned with *concepts*, while modelling is concerned with *types*. A concept may be associated with zero, one, or more types, and a type is always associated with a single concept. Concepts are defined, while types are specified. The distinction is that a concept states the condition that objects must meet to become classified (identified) as an instance of its associated concept/class. A type, on the other hand, contains a list of attributes that are used to describe the objects that become instances of the associated concept/class. The distinction can be illustrated as follows:

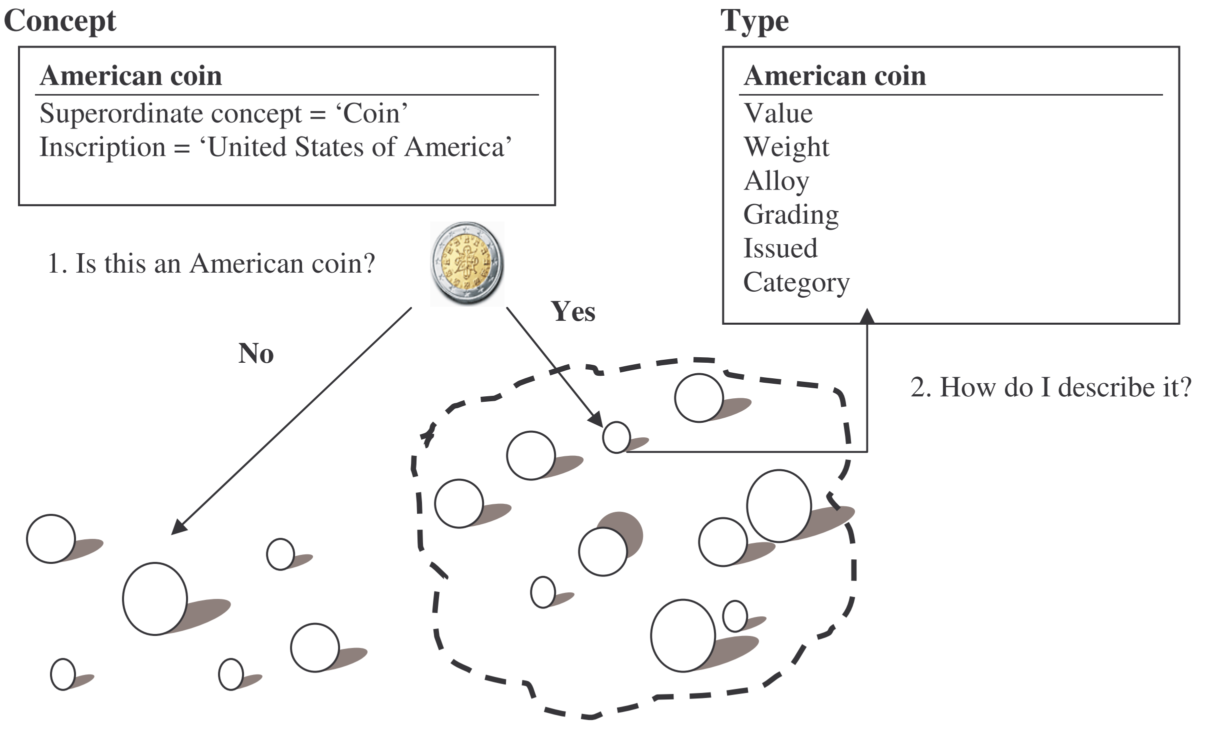


Figure 4.11: Classification versus modelling.

The concept helps us to determine which coins are American. The type tells us how to describe those coins we identify as American coins. The concept contains *defining* properties. The type contains *describing* properties. Maybe the two can be combined and represented as a single construct depicted in figure 4.12:

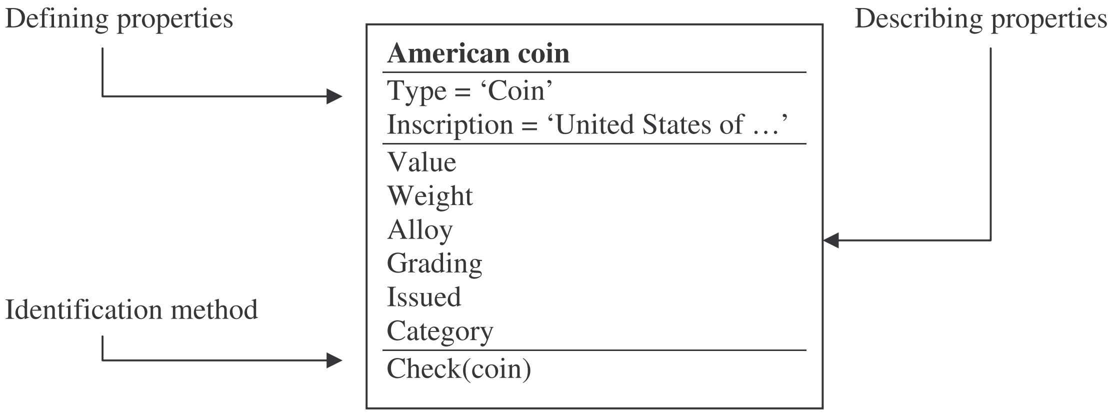


Figure 4.12: Extended type construct.

4.3.3 Benefits of classification.

In Moody and Shanks (1998), the authors present a framework to evaluate the quality of entity relationship models based on seven quality factors: *correctness, completeness, simplicity, flexibility, integration, understandability,* and *implementability*. Of these, classification may have a positive impact on completeness, integration, and understandability.

Completeness refers to whether the data model contains all information required to meet the user requirements. In this respect classification may contribute with a *complete* and *logically consistent* vocabulary. The problem of not having a vocabulary is addressed by Moody and Shanks (1998):

> "In principle, completeness can be checked by checking that each user requirement is represented somewhere in the model, and that each element of the model corresponds to a user requirement (Batini *et al*, 1992). However, the major difficulty with checking completeness is that there is no external source of user requirements – they exist only in people's minds. As a result, completeness can only be evaluated with close participation of business users", (Moody and Shanks 1998, p. 102).

The vocabulary will be *complete* when it contains the concepts necessary to express the user requirements, and it will be *logically consistent* when all concepts are defined with intensional or extensional definitions. However, in practical terms, it is generally advised that the need for rigour in definitions must be balanced with the requirement for it to be practical and useable. Therefore, it may be unrealistic to expect a vocabulary in which each and every term is fully consistent with every other term.

Another aspect of completeness is the representation of integrity constraints. In this respect, classification adds membership conditions that may be used to control the quality of data input.
Membership conditions may be implemented as administrative routines, as automatic procedures, or as a combination of both. As an example, see the identification method in figure 4.12.

Integration is another quality factor which is defined as the consistency of the data model with the rest of the organization's data. The authors mention three aspects to integration: *data sharing-reuse, consistent definitions*, and *corporate view of data*, all of which require explicitly defined membership conditions. To successfully share, or reuse data sources, it is necessary to know and to compare the membership conditions associated with each class. The reason for this is that objects may be described the same way, but still be of different kinds, or, alternatively, be described in different ways, while still being of the same kind. Consistent definitions facilitate comparability and consolidation of data across applications. Again, the very notion of consistent definitions is the result of classification, and also one of its senses. Lastly, corporate view of data suggests an approach where data should be defined in a way that is useable across the organization, in order to avoid narrow, and application specific definitions. This also implies the need for a corporate-wide vocabulary.

A third quality factor is *understandability*, which is defined as the ease with which the concepts and structures in the data model can be understood. Business users must be able to understand the model in order to verify that it is a complete and accurate representation of their requirements. To see how membership conditions may contribute with respect to understandability, verification and validation, consider the examples in figure 2.8 and 2.9 at page 59.

4.4. Conceptual modelling – practical aspects

In this section, conceptual modelling is presented in the six steps shown in figure 4.13. The number of steps is not important. It could have been 4 or it could have been16, but in this context, six steps are found sufficient to address the necessary elements of classification.
The modelling process is presented as a sequence of steps, but in a practical setting, the steps may be conducted in any order, as long as there are some initial user requirements to start from. When several persons are involved, the process may even be run concurrently. Some may collect user requirements, while others may work with the identification and definition of concepts.

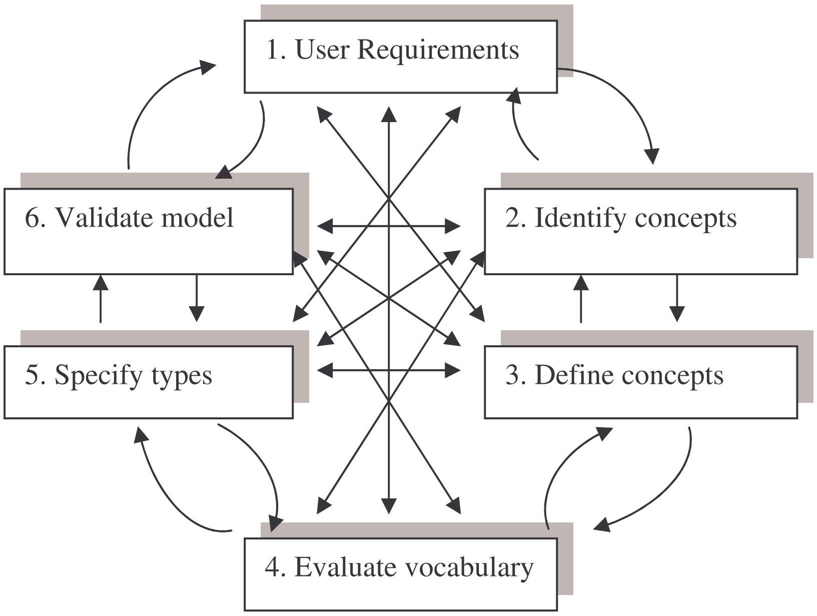


Figure 4.13: The baseball model. Adopted and adapted from Coad and Nicola (1993).

Lastly, conceptual modelling is normally considered as a single stage in a micro life cycle, commonly referred to as *the database application lifecycle*, which, in turn, is part of a larger, macro life cycle which covers the life cycle of the complete information system. For further details on the interplay between the micro and macro life cycles, see Connolly and Begg (2002), p. 270, and Elmasri and Navathe (2000), p. 530.

4.4.1 Collect user requirements.

Ideally, classification starts with the user requirements capture. Depending on the particular methodology used, the requirements specification document may be required to follow certain standards with respect to its form and content. Normally it will contain statements about the overall purpose of the system, its scope and boundaries. It may also contain mission objectives, i.e., statements about the particular tasks the system is expected to support, along with particular user requirements for the new system.

The content of the requirement specification document may be collected by different fact-finding techniques. Figure 4.14 shows five of the most commonly used techniques:

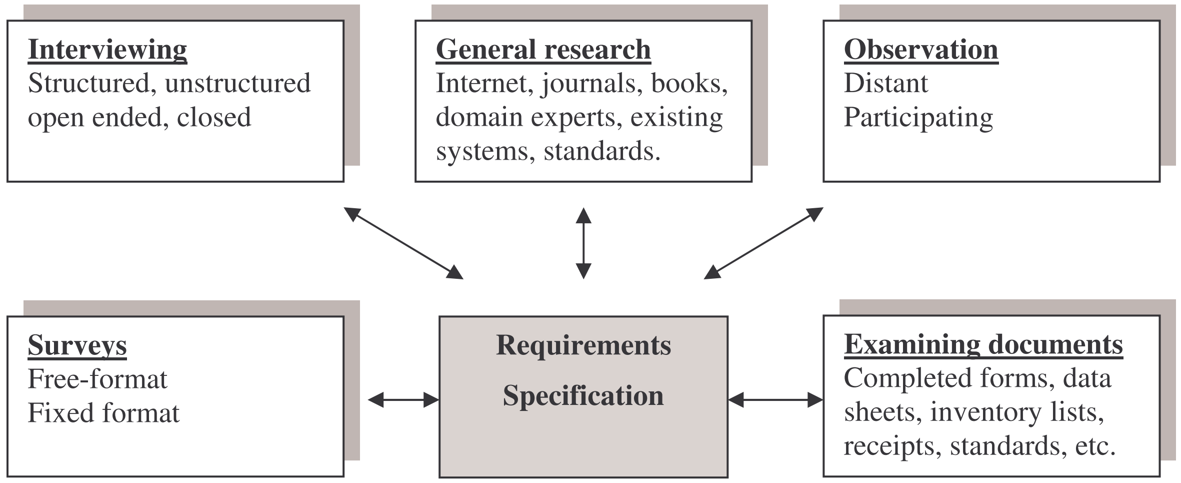


Figure 4.14: Fact-finding techniques commonly used to collect user requirements.

During requirements capture, information is collected, analyzed and arranged according to the particular guidelines of the chosen methodology. Although none of the above techniques focus on the definition of concepts and terms, concepts are still easy to identify from the resulting specifications. Take, for example, the following list of mission objectives taken from Connolly and Begg (2002):

To maintain (enter, update, and delete) data on branches.
To maintain (enter, update, and delete) data on staff.
To perform searches on branches.
To perform searches on staff.
To report on branches.
To report on staff.

With classification in mind, the terms *branch* and *staff* stand out as typical terms or concepts that need to be further defined. In addition, searches, as those stated in line 3 and 4, require search conditions, which in turn may indicate predicates that may suggest subordinate concepts, such as *full time*, versus *part time staff*, and *technical* versus *administrative staff*. So concepts should not be too hard to identify. However, if one needs further guidelines to identify concepts, some concept analysis approaches are dealt with below.

4.4.2 Identify concepts.

The literature contains many different approaches to find objects and concepts, some of which will be demonstrated in this subsection. What they all have in common is that they are variations of concept analysis techniques. Booch (1991) lists four different techniques: object-oriented analysis, domain analysis, text-analysis, and structured analysis.

Object- oriented analysis focuses on the identification of classes and objects that form the vocabulary of the problem domain. The identification process is generally guided by category lists, such as the ones proposed by Schlaer and Mellor (1988), Ross (1987) cited from Booch (1991), and Coad and Yourdon (1991), respectively:

Category	Examples
Tangible things	Cars, telemetry data, pressure sensors
Roles	Mother, teacher, politician
Events	Landing, interupt, request
Interactions	Loan, meeting, intersection
Specifications	Types, categories, models

Table 4.1: Schlaer and Mellor's list to identify classes and objects.

Category	Examples
People	Humans who carry out some function.
Places	Area set aside for people or things.
Things	Physical objects, or groups of objects, that are tangible
Organizations	Formally organized collections of people, resources, facilities, and capabilities having a defined mission, whose existence is largely independent of individuals.
Concepts	Principles or ideas not tangible per se; used to organize or keep track of business activities and/or communications.
Events	Things that happen, usually to something else at a giveren date and time, or as steps in an ordered sequence.

Table 4.2: Ross' list to identify classes and objects.

Category	Examples
Structure	“Kind of” and “part of” relationships.
Other systems	External systems with which the application interacts.
Devices	Devices with which the application interacts.
Events remembered	A historical event that must be recorded.
Roles played	The different roles users play in interacting with the application..
Locations	Physical locations, offices and sites important to the application.
Organizational units	Groups to which users belong.

Table 4.3: Coad and Yourdon’s list to identify classes and objects.

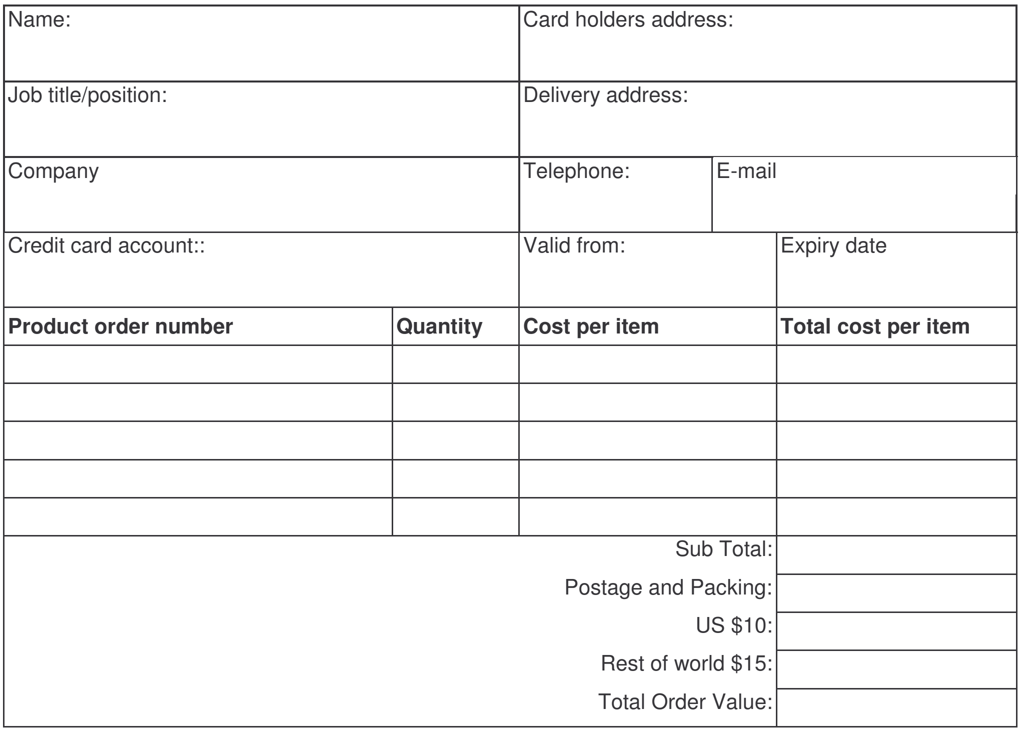
A second approach, *domain analysis*, is defined as an attempt to identify the objects, operations, and relationships that domain experts perceive to be important about the domain. A domain expert is a person who is intimately familiar with all the elements of a particular problem and one who speaks the vocabulary of the problem domain.

A third approach is *text analysis* of informal problem descriptions. By this method, candidate classes are identified by the nouns in the text, and candidate operations are identified by the verbs. In the example text below, the nouns that suggest candidate concepts have been emphasized.

> When entering new **orders** the system must generate a unique **order number** and assign the correct **date** to the **order**. Based on the **customer number** the system must output **name**, **category** and **customers address**. Then, according to the **customer’s specifications**, one or more **order-lines** must be entered with **product number** and **quantity** per product. Based on the **product number** and the **quantity** entered, the system must output the **name** of the **product**, compute the **total amount** for each **order-line,** and reduce **quantity at stock** accordingly.

A fourth approach is *structured analysis*, which is an analysis technique that produces context diagrams, data dictionaries, and data flow diagrams. Candidate concepts may be identified by inspecting the data dictionary elements, and by studying external entities, data stores, control stores, data flows and control flows.

A fifth approach is called *fact-based analysis*, and focuses on the analysis of structured documents, (Edmond 1992). Structured documents, such as forms, screens, reports, and data sheets, represent standardized and well thought out data inputs and outputs to manual and/or computerized information systems. Usually such documents exhibit well organized information structures that may be used to identify primitive concepts as well as composite concepts. For instance, in figure 4.15, one finds primitive concepts such as *name, job title*, *company*, etc., organized in clusters that represent composite concepts such as *customers*, *accounts*, and *order lines*. In addition, some of the primitive concepts, such as *company* and *product number* indicate relationships to other composite concepts, such as *companies* and *products*.



Name:		Card holders address:	
Job title/position:		Delivery address:	
Company		Telephone:	E-mail
Credit card account::		Valid from:	Expiry date
Product order number	**Quantity**	**Cost per item**	**Total cost per item**
		Sub Total:	
		Postage and Packing:	
		US $10:	
		Rest of world $15:	
		Total Order Value:	

Figure 4.15: An order form containing primitive concepts arranged in clusters.

According to Kim and March (1995) there are essentially two types of data modelling formalisms: entity-attribute-relationship (EAR) models and object-relationship (OR) models. Both types of models distinguish between primitive and composite objects. EAR-models use the terms *entity* and *relationship* to denote complex objects, and the terms *attribute* and *value-set* to denote primitive objects. Similarly, OR-models use the terms *NOLOT* and *Role* to denote complex, non-lexical objects, and *LOT* to denote primitive, lexical objects. Complex objects are objects that need to be further described by more primitive objects. The notions of complex and primitive objects are relative terms. In one context an object may be conceived of as a complex object, while in another context it may be conceived of as a primitive concept. It all depends on user requirements and the nature of the application. Corresponding to complex and primitive objects, there are complex and primitive concepts. A complex concept represents complex objects, and a primitive concept represents primitive objects.
As concepts are identified, they need to be further analyzed to be correctly defined. This is described in the next subsection.

4.4.3 Define the concepts

For concepts to be properly defined, they must be analyzed and arranged in a concept system. A concept system is an arrangement of concepts that are related by generic, partitive, and/or associative relationships. The general idea is that the final definitions shall be coordinated so as to reflect the concept system by indicating the connection to other concepts or the delimitation that distinguish one concept from another. The analysis consists of five interrelated tasks:

1. Separating concepts that need to be defined from concepts considered to be basic or familiar.
2. Analyzing the intension and extension of each concept, starting with the most superordinate concepts.
3. Determining the relation and position of each concept within the concept system.
4. Formulating and evaluating definitions for the concepts based on the concept relations.
5. Attributing designations to each concept.

4.4.3.1 Separate concepts.

A *student name* is a *name*, which is a *string* (at least for those of us who work with databases), which is a *data type, which is …*. A definition starts a process that can go on ad infinitum. For practical reasons, some terms must be allowed to remain undefined, in order to terminate an otherwise infinite regress.

Some terms may go undefined because they are already defined by the modelling formalism that is to be used. For instance, the ER-model defines terms like *entity*, *entity set*, *relationship*, *relationship set*, *domain*, and *attribute*. Other terms, like *number*, *string*, and *data type* are defined by the target DBMS. In addition, within a given context, some concepts are so basic or familiar that explicit definitions are unneeded. Hence, unless the term *Student name* has a very special meaning, there may be little to gain from defining it.

Therefore, as a general advice, to be fully defined, terms must "deserve" their definitions. For a term to deserve its definition, there must be a reasonable chance for the term to be *misinterpreted*, and the misinterpretation must have a *consequence*. Terms that are not defined on the other hand, should all together be listed as undefined terms. Consequently, the output from this task is a list of terms that are to remain undefined, and a list of terms to be defined.

4.4.3.2 Analyze the intension and extension of each concept.

Depending on the rigour with which the remaining analysis is conducted, dictionaries and thesauruses may be consulted, and the concepts analyzed by considering possible supertypes, subtypes, partitions and associations. For each concept, the following aspects must be further explored:

1. inclusion criteria,
2. typical instances and borderline cases,
3. non-essential and/or optional characteristics of the instances,
4. the source or informant from which the definition is obtained.

4.4.3.3 Determine the position of each concept within the concept system.

As the concepts are analyzed, the concept system will need to be sketched and tentative definitions drafted, evaluated, and reworked. The final definitions should reflect their position in the concept system by including references to related concepts.

For example, the definition below indicates how the term *Teacher* is related to *Person* by a generic relationship and to *Course* by an associative relationship. This suggests an initial concept system as depicted in figure 4.16.

A *teacher* is a *person* who teaches at least one *course*.

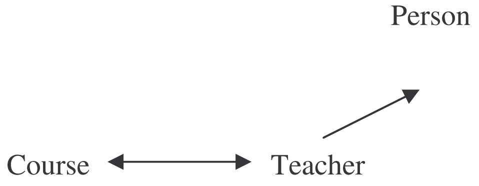


Figure 4.16: Simple definition and associated concept system.

On closer evaluation of the definition we may come to learn that there are staff members and students, and that among the staff members there may be teachers, researchers, and administrative personnel. This will cause the concept system and its definitions to be reworked. By repeating the process the concept system and its vocabulary is gradually expanded.

4.4.3.4 Formulating and evaluating definitions.

Each definition must be represented by a terminological entry as described in section 4.3.1. In addition, the following aspects must be further explored:

- consequences for misclassification
- identification procedures and any special administrative and/or technical requirements

Each definition, or terminological entry, must be complemented by a note, in which at least the following extra information is documented:

1. A justification for the membership condition chosen for the concept.
2. An analysis of possible consequences from misclassification.
3. An identification procedure to ensure consistent application of the concept.
4. Administrative and/or technical requirements needed for the identification procedure.
5. The source - the informant or reference from which the definition is obtained.

For further details, see example on next page.

4.4.3.5 Attributing designations

Finally, a term to denote the concept must be decided for. According to the ISO 704:2000 Standard, established, and widely used designations should be chosen, even if they are poorly formed or poorly motivated. If several designations exist for a single concept, then the one that satisfies the largest number of principles listed below should be selected:

1. Transparency – the term's meaning can be at least partly inferred without a definition.
2. Consistency – the term must integrate with or be consistent with the concept system.
3. Appropriateness – adherence to familiar, established patterns of meaning.
4. Linguistic economy – terms should be as brief and concise as possible.
5. Derivability – productive terms with many derivatives should be preferred.
6. Linguistic correctness – terms that conform to traditional language norms.
7. Preference for native language – native expressions should be preferred for direct loans.

Member:

A member is a person who has paid the membership fee for the current membership period.

Justification: Political parties receive public financial support from the central political authorities every year. The amount received is computed from the current year's total member count. To count as a member, membership fee for the current period must have been paid. The total member count must be correctly reported with the application for next year's support.

Consequences: Failing to correctly identify this year's members when new members are recorded, or existing members are updated, an incorrect member count may be reported. If the report is incorrect, the party risk to return this year's amount, and to be cut off from any financial support in the future. In addition, the party may loose its credibility in the public and its popularity may decline.

Identification procedure: Members pay the membership fee to a bank account. Every Friday the bank reports the last weeks transactions, including name and address of the payer, amount paid, and date of payment. If the amount is valid, the member is looked up and the transaction number is updated from last year's number to this year's number. In addition, the date of payment is updated from last year's date to this year's date. If the member cannot be found, a new member is inserted, with member number, name, address, transaction number and date of payment.

Administrative requirements: The transaction drafts received from the bank must be kept in a separate folder for 5 years, in case of external audits by the auditor-General. It is the secretary's responsibility to update the data base on a regular basis and to safely store the drafts.

Technical requirements: Physical folder in a fire proof cabinet.

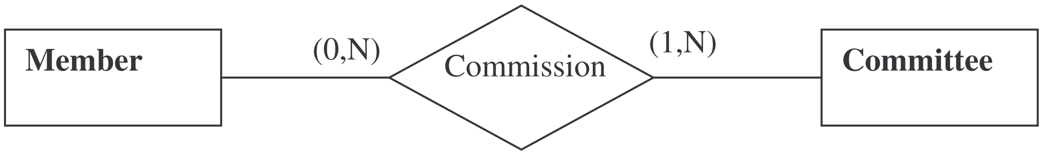
Source: Interview with local party leader, and information from the auditor-General.

Figure 4.17: Example form for the definition and extra information about a term.

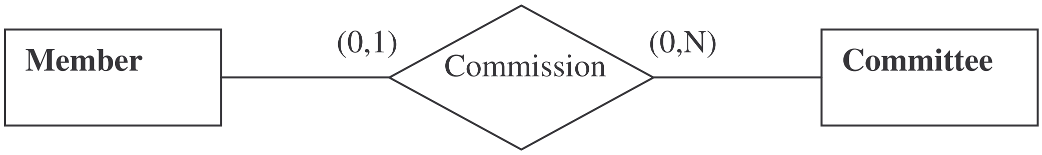
As the example demonstrates, the membership condition is that the members have paid the membership fee. However, the *defining* properties which will be used to document that the membership condition has been met are *transaction number* and a legal *date of payment*. These will only be updated or recorded when information about the transactions has been received from the bank.

To the extent that relationships between concepts are being named, membership conditions for relationships should be included in the vocabulary. Terms that denote relationships may generally be defined by a single format as demonstrated below:

R is a relationship between **X** and **Y**. A **x** *may* | *must* be related to *one* | *many* | *n* **y**. A **y** *may* | *must* be related to *one* | *many* | *n* **x**.



A commission is a relationship between Member and Committee. A member may be related to many committies. A committy must be related to many members.



A commission is a relationship between Member and Committee. A member may be related to one committie. A committy may be related to many members.

Figure 4.18: Definition of terms denoting relationships.

4.4.4 Evaluate the vocabulary.

The vocabulary may be evaluated for its *completeness*, *logical consistency*, *understandability* and *parsimony*.

The vocabulary will be *complete* when it contains the concepts necessary to express the user requirements. Hence, the evaluation can be conducted by reviewing the requirements specification to check if the key terms are contained in the vocabulary.

The vocabulary will be *logically consistent* when all concepts are defined with intensional or extensional definitions. Another way of formulating this is to say that a vocabulary is logically consistent if its concept system can be generated from its definitions.

The use of intensional and extensional definitions requires that every term referred to in a definition must be defined, or at least contained, elsewhere in the vocabulary. This requirement makes it possible to conduct systematic and formal checks by at least two approaches:

1. Generate a concept system from the vocabulary and compare it with the original concept system.
2. Control each definition to check that it only contains terms defined elsewhere in the vocabulary.

However, logically consistency is a formal property. Definitions that are logically consistent may still not be meaningful or coherent in any sense. Hence, it is important to evaluate the vocabulary for its understandability as well.
The vocabulary is *understandable* if the users, analysts, designers and other stakeholders understand, and agree to the definitions and their associated notes concerning the justification, consequences, identification procedures, administrative and technical requirements and the authority of the source. According to Moody and Shanks (1998), understandability may be evaluated by means of user reviews, scenario analysis, and application developer reviews.

Finally, the vocabulary may be evaluated for its *parsimony and elegance*. A vocabulary based on simple definitions and only those concepts that are required for the solution of the problem is preferred to more complex organizations.

4.4.5 Specify types and complete the conceptual model.

The specification of types is to a large extent determined by the selected modelling formalism. If the Entity-Relationship model is being used, then at least value sets, entity types, and relationship types need to be explicitly specified. Other formalisms may have other requirements. Below are some guidelines which are especially tailored towards the use of EAR-formalisms, such as the Entity-Relationship model and the Enhanced Entity-Relationship model.

1. Which attribute(s) will be used to represent *the individual identity* of each entity in the associated entity set? If a decision has been made, then add the attribute(s) to the type, otherwise, ask again during logical design.
2. Which attribute(s) will be used to represent *the typological identity* of each entity in the associated entity set? Check the vocabulary! If the attribute(s) serve(s) *descriptive* purposes as well as *defining* purposes, like *transaction number* and *date of payment* in figure 4.19, then add the attribute(s) to the type. If, however, the attribute(s) serve a defining purpose only, they may be treated as constants and represented as a class property to avoid redundancy.
3. Which attribute(s) will be used to *describe* each entity in the entity set? Check the requirements specification and the vocabulary for relevant concepts and add them to the type.
4. How will the membership condition be *controlled*? Check the vocabulary and add a method to the type if the formalism allows it. At a later stage, write an algorithm for the method.
5. How should the model constructs be *named*? Use the terms in the vocabulary to name the value sets, entity types, attributes, relationship types, and roles. If new terms are introduced to name constructs in the model, then add the new terms to the vocabulary.
6. For each generic, or partitive relationship that is mirrored in the conceptual model, attributes corresponding to the *dimension criteria* should be added to the type. A dimension criterion is a predicate that controls the subdivision of a concept into subordinate concepts.
7. The conceptual model may mirror the concept system with respect to its concepts and its generic, partitive and associative relationships. On the other hand, the conceptual model may deviate from the concept system structures. For instance, a generic hierarchy in the concept system may be conflated in the conceptual model, and the subordinate terms in the vocabulary may be used as role-names in the conceptual model as shown in figure 4.19, and 4.20. Any such deviations in the conceptual model should be explicitly justified and documented as part of the model specification.

The concept system in figure 4.19 shows two generic hierarchies, *Person* and *Paper*, and their respective dimension criterions, *Role* and *Status*. An associative relationship is shown between *Session Chairman* and *Session* and a partitive hierarchy is shown between *Session* and *Accepted Paper*.

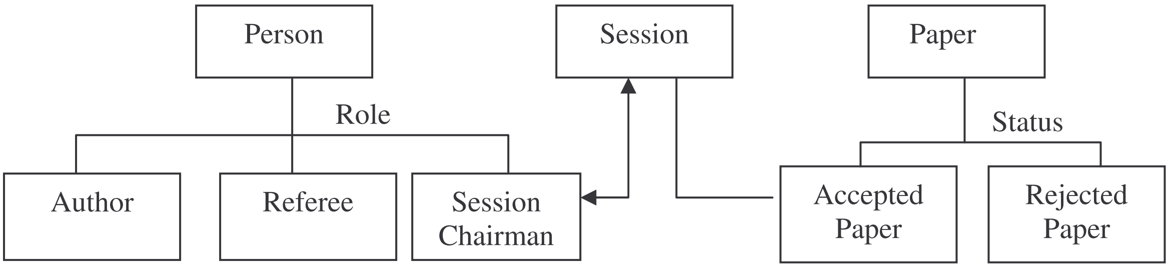


Figure 4.19: Example concept system.

However, the conceptual structures in the vocabulary need not be directly mapped onto the types in the conceptual model. For various reasons, it may be decided to represent the Person hierarchy with a single type, and to use role names to represent authors, session chairman, and referees. Similarly, it may be decided to exclude rejected papers all together in the conceptual model. In this respect, the vocabulary may provide a basis to discuss and justify such design decisions.

The refereeing type in the conceptual model illustrates the iterative aspect of the classification process. New types may be introduced in the conceptual model, for instance to represent a n:m relationship. In such cases, the new concept must be included in the vocabulary.

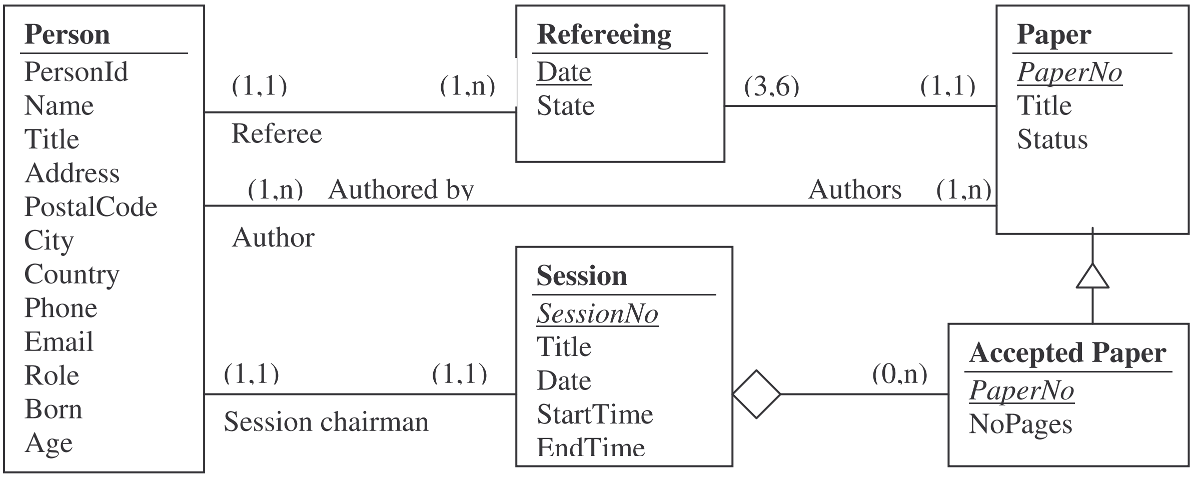


Figure 4.20: Example conceptual model. See notation in appendix G.

4.4.6 Validate the model according to the vocabulary.

The final task serves two purposes: First, to validate the model according to the vocabulary, in order to ensure that all terms that are used in the conceptual model are contained in the vocabulary. As demonstrated in figure 4.20, new terms, such as *Refereeing*, *Authored by* and *Authors* must be defined in the vocabulary.

Second, any deviations in the conceptual model must be justified and documented. In figure 4.20, explanations are required as to why the *Person* generic relationship is conflated, and the *Paper* generic relationship is simplified by discarding *Rejected papers* from the model.

5.0 Empirical study of interpretation tasks

5.1 Introduction.

One of the implications from the concept analysis in chapter 2 is that classification may have an effect on the interpretation and the design of conceptual models. This chapter describes an experimental research design that was conducted to empirically test the effect of classification on *interpretation* tasks. The basic idea behind the experiment is to study how knowledge of membership conditions affects people's interpretation and confidence when they are presented to simple conceptual models such as the models in figure 2.8 and figure 2.9 in chapter 2.

5.2 Related work

Various empirical studies have been conducted on the usability of data models with respect to model construction and model validation, but none of the studies are concerned with membership conditions or related notions. Still, these studies are interesting in that they suggest various approaches to test data models and data modelling.

Studies by Batra, Hoffer, and Bostrom (1990), Batra and Antony (1994), compare designer performance in modelling tasks using the relational model and ER-model as independent variables, and modelling correctness as the dependent variable. The major conclusions from these studies are that ER and EER-models are superior to the relational model when it comes to correctness in modelling relationships, and that the predominant errors in ER modelling were incorrect representation of connectivity of relationships. These results were further investigated in several experiments by Siau, Wand, and Benbasat (1995), (1997), and by Dunn and Grabski (2001). In their first study, Siau, Wand, and Benbasat (1995), investigated the use of optional and mandatory relationships by expert users. Having the subjects choose the correct participation constraint for both familiar and unfamiliar relationships, they found a tendency by expert users to prefer optional over mandatory relationships. In their next study, Siau, Wand, and Benbasat (1997) investigated whether expert users focused on the structural constraint, or the underlying semantics represented by the names of the relationship and the participating entity types. The results of the study indicated that the subjects ignored the underlying semantics and based their judgments almost exclusively on the structural constraints.

To gain further insight on the results from Siau, Wand, and Benbasats' studies, Dunn and Grabski (2001) extended the studies by framing the questions in such a way that the subjects were forced to apply either syntactic or semantic understanding to the interpretation tasks. In addition, they considered the effect of a greater information load by having the subjects consider relationships from the greater context of full ER-diagrams. The results indicate that when asked *what* the ER diagram portrays, the participants focused on the structural constraints and exhibited syntactic understanding. When asked to identify *appropriate* participation, the participants exhibited semantic understanding. It was also shown that information load affected the participants' semantic understanding of relationship participation.

The current study is heavily influenced by the three studies above. Having the notion of membership conditions in mind, it seemed a good idea to extend the studies, by adding membership conditions to the interpretation tasks. However, for reasons described in section 5.3, some changes had to be made to the instruments.

In terms of the process model for information requirement determination, described in chapter 1, Kim and March (1995) study the effects of different data modelling formalisms on modelling and validation tasks. They suggest that future research should examine the effects of alternative modelling formalisms on the discovery task.

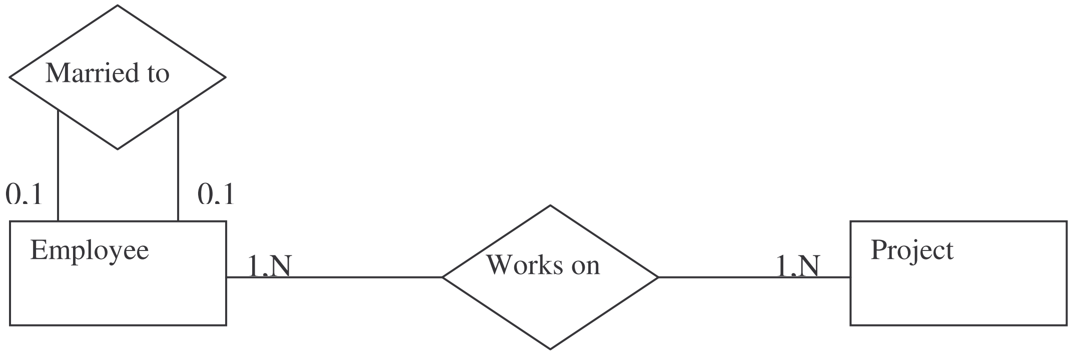
Although the current study does not involve a comparison of different modelling formalisms, it does address both the discovery phase and the validation phase, by considering the effects that knowledge of membership conditions from the discovery phase may have on interpretation tasks.

5.3 Research question

According to Kim and March (1995), and Siau, Wand and Benbasat (1997), there are two task categories in information modelling – model interpretation and model construction. The research question and the experimental design in this study are motivated by previous studies on how experts and novices *interpret* relationships in ER-diagrams.

The initial idea was to add membership conditions to the interpretation tasks used in Siau, Wand, and Benbasat (1995), and to let the participants choose the connectivity of each relationship. A comparison would then be made between the results of the current study and those of the previous study. However, it soon became clear that for the current study, it would be better to use tasks from a single application domain, rather than individual, generic tasks.

The reason for this is that the membership condition for an entity type may not be equally informative for all the relationships in which the entity type participates. This can be illustrated by an example:



An employee is a person who works on at least one project at any time.

A project is a complex work task which involves the joint effort of several employees.

Figure 5.1: Informative and non-informative membership conditions.

While the membership condition of Employee is informative with respect to the relationship *Works on*, it does not give any clues at all when it comes to the *Married to* relationship. Based on the membership conditions for Employee and Project, one can be quite confident that the connectivity of the *Works on* relationship is correctly displayed in the ER-diagram. When it comes to the *Married to* relationship however, we can only assume that the connectivity is correctly displayed, based on common sense. The assumption would be correct if the *Married to* relationship is meant to represent only current marriages between employees. It would be incorrect if the relationship is meant to include past marriages between employees as well. Hence, when interpreting relationships, some interpretations may be based on membership conditions, while others can only be based on common sense. In order to have both informative and non-informative membership conditions reflected among the interpretation tasks, all tasks were selected from a full ER-diagram. In addition, a full ER-diagram will take care of any effects from participation in multiple relationships as reported in Dunn and Grabski's study (2001).

At this point, it should be noticed that in a fully specified conceptual model, there will be membership conditions for relationship types as well as for entity types, since both constructs represent classes.

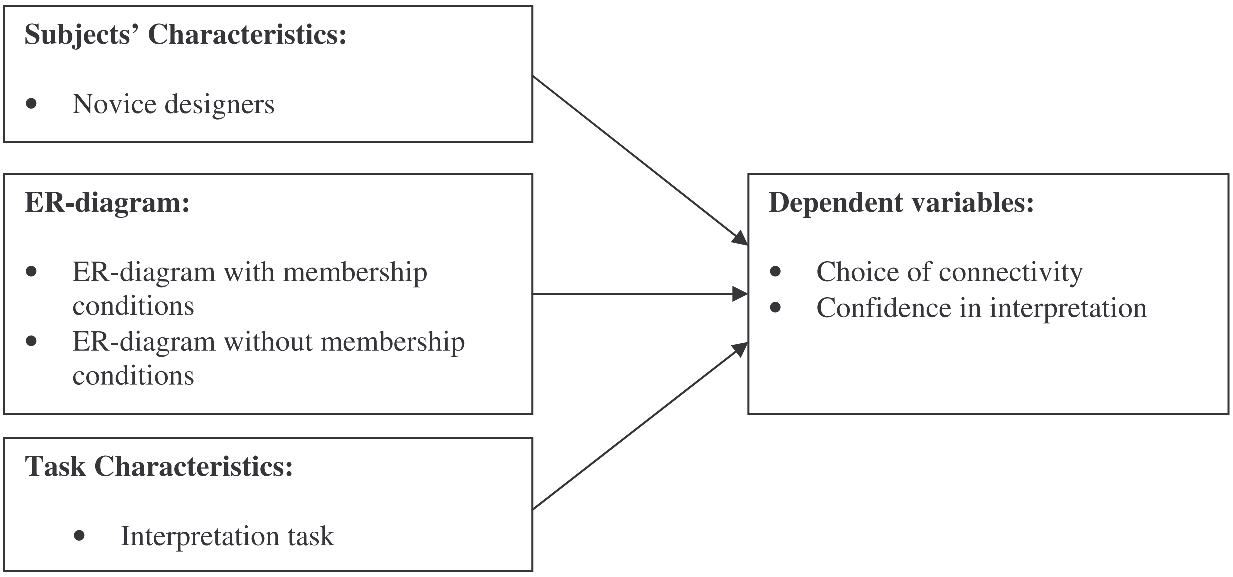
It is only for the purpose of measuring the effect of membership conditions that membership conditions for relationship types have not been explicitly stated for each interpretation task. In other words, it is assumed that knowledge of the membership conditions for the entity types will be of help to interpret and validate relationships.

Accordingly, the research question for this study will be addressed by measuring the effect of membership conditions on:

- The participants' interpretation of relationships in a conceptual model.
- The participants' confidence in the interpretations.

5.4. Experimental design and framework

The experiment consisted of two groups of undergraduate information science students, taking an introductory course in data modelling. The first group was exposed to binary relationships from a full ER-diagram, with membership conditions included for the participating entity types. The second group was exposed to the same set of binary relationships, but without membership conditions. The research framework is illustrated in figure 5.2, and the questionnaires can be found in appendix A.



Subjects' Characteristics:

- Novice designers

ER-diagram:

- ER-diagram with membership conditions
- ER-diagram without membership conditions

Task Characteristics:

- Interpretation task

Dependent variables:

- Choice of connectivity
- Confidence in interpretation

Figure 5.2: Experimental framework adopted from Siau, Wand, and Benbasat (1997).

In preparing the experiment, it was ascertained that the subjects had no prior knowledge of membership conditions. Hence, it could not be anticipated that the subjects would be able to fully take advantage of the existence or absence of the membership conditions during the first experiment. Therefore it was decided to lecture the subjects on the notion of membership conditions, and then to repeat the experiment a second time.

In the first two-hour lecture, the students were trained in ER-modelling, with a special focus on the interpretation of relationships, with respect to connectivity constraints, and the notation used for the experimental tasks. Nothing was said about membership conditions. After a 20 minutes training session, the questionnaires were handed out, and the experiment was run for the remaining 20 minutes. After a 15 minutes break, the students received a 45 minutes lecture on concepts, membership conditions, concept definitions, and the distinction between classes and types. The day after, the second experiment was run for 20 minutes, followed by a lecture on how to identify, document, and include membership conditions in conceptual modelling tasks.

5.4.1 Independent variable

The independent variable in this study is the ER-diagram. One group of subjects received the full set of binary relationships extracted from the ER-diagram without membership conditions, as illustrated in figure 2.8, while the other group receive the same set of relationships with membership conditions included, as illustrated in figure 2.9. The full set of binary relationships are arranged in the same order as if the ER-diagram would be read from left to right, top to bottom. In addition, the full ER-diagram, without connectivity constraints and membership conditions were printed on page 1 of each questionnaire. The questionnaires were sorted such that every second questionnaire contained interpretation tasks with membership conditions. The subjects were then randomly assigned to the two treatments by handing out the questionnaires one by one in the order the subjects were seated.

5.4.2 Dependent variables

There are two dependent variables in this study: choice of connectivity, and confidence level. The choice of connectivity is captured by means of multiple-choice questions. For each task, the students are asked to choose one from among four connectivity options, represented by 0..1, 0..N, 1..1, and 1..N. To capture the confidence in interpretation, the subjects choose a value from a 7-point scale, where 1 indicate no confidence, and 7, absolute confidence.

5.4.3 Task characteristics

The two ER-diagrams in the study are based on two design tasks created and used in Shoval and Shiran (1997). In their paper, the tasks are used to compare EER and OO data models from a designer perspective. In the current study, the two tasks are used for model interpretation. The reasons for choosing the tasks are that they a) have been successfully used in an experiment, b) are well documented with both a narrative description and a solution diagram for each task, c) are similar in size and complexity, and d) the same tasks may be used in a subsequent study to investigate the effects of membership conditions on model construction.

5.4.4 Subjects

The subjects in the study were information science undergraduate students taking introductory courses in information and communication technology, and systems analysis and design. The two experiments were conducted as part of two guest lectures in the systems analysis and design course. At the time of the experiment, the students had received 3 two-hour lectures in ER-modelling and completed two group-exercises in data modelling, guided by a teaching assistant.

In the first experiment, a total of 33 students attended the lecture. The following day, 5 of the subjects were unable to attend, leaving 28 subjects for the second experiment. In each experiment, the subjects were randomly assigned to two treatments. One group received ER-diagrams without membership conditions, the other group received ER-diagrams with membership conditions. The number of subjects and the number of observations for each group and each experiment are shown in table 5.1 and 5.2.

	No. of subjects	No. of questions per subject	No. of observations
Without membership conditions	16	20	320
With membership conditions	17	20	340

Table 5.1: Number of subjects for each treatment in experiment 1.

	No. of subjects	No. of questions per subject	No. of observations
Without membership conditions	14	24	336
With membership conditions	14	24	336

Table 5.2: Number of subjects for each treatment in experiment 2.

On the last page of each questionnaire, demographic information was collected from each subject. In the first experiment there were 19 male and 14 female students, and in the second, 15 males and 13 females. Their average age was 24 years. The subjects' expertise in conceptual modelling was measured on a scale from 1 to 5, ranging from no experience, or no confidence, to highly experienced, or highly confident.

Demographics	**Experiment 1**		**Experiment 2**	
	Mean	**Std.Dev**	**Mean**	**Std.Dev**
Data Modelling Experience	1.93	1,07	1,86	1,13
ER Model Experience	1,89	1,01	1,82	0,96
ER Model Confidence	2,56	0,89	2,55	0,86
ER Model Syntax Familiarity	2,78	1,22	2,86	1,04
Relevant practice in years	0,54	2,04	1,07	2,96

Table 5.3: Subjects' self-reported expertise.

As reflected in table 5.3, the subjects can be considered as novices with respect to conceptual modelling. They have almost no experience in data modelling in general, and almost no experience in using the ER-model. However, based on previous lectures, they feel moderately confident with the ER-model and the ER-notation used in the experiment. The difference in the mean score on the practice variable is due to a single subject reporting 10 years of practice. The median score for this variable is 0,00 in both experiments.

5.4.5 Hypotheses

It is expected that knowledge of membership conditions will have an effect on the choice of interpretation, and that the subjects will feel more confident with respect to their choices, when membership conditions are included. Since the control group can only make qualified guesses about the connectivity constraints, the responses from this group cannot be directly compared to those of the treatment group. Instead, the responses from the control group are used as a benchmark, and an independent samples t-test is used to measure the difference between the mean scores of the two groups with respect to the benchmark.

Similarly, an independent samples t-test is used to measure the difference between the mean confidence scores in the two groups.
Since some of the interpretation tasks in the treatment group contain non-informing membership conditions, it is further expected that the differences are larger when only interpretation tasks with informative membership conditions are compared. Finally it is expected that there will be no significant differences between the two groups with respect to interpretation tasks when only non-informative membership conditions are considered.

This leads to the specification of 6 null hypotheses:

$H1_0$: There is no difference in the choice of interpretation between the treatment group and the control group.

$H2_0$: There is no difference in the confidence of interpretation between the treatment group and the control group.

$H3_0$: There is no difference in the choice of interpretation between the treatment group and the control group when only interpretation tasks with informative membership conditions are considered.

$H4_0$: There is no difference in the confidence of interpretation between the treatment group and the control group when only interpretation tasks with informative membership conditions are considered.

$H5_0$: There is a difference in the choice of interpretation between the treatment group and the control group when only interpretation tasks with non-informative membership conditions are considered.

$H6_0$: There is a difference in the confidence of interpretation between the treatment group and the control group when only interpretation tasks with non-informative membership conditions are considered.

5.4.6 Experimental results

Experiment 1

Table 5.4 depicts the responses from the control group in experiment 1. For each interpretation task, the most frequent choices are coloured grey. These choices are used as a benchmark, in order to measure how membership conditions affect the choices made by the treatment group.

Interpretation tasks	**Respons counts**				**Total**
	0..1	**0..N**	**1..1**	**1..N**	
1. Supplier-agreement-dept		13		3	16
2. Dept-agreement-supplier		15		1	16
3. Dept-belong_to-employee	1	4		11	16
4. Employee-belong_to-dept		3	8	5	16
5. Dept-managed_by-manager	3	2	7	4	16
6. Manager-managed_by-dept	3	2	5	5	15
7. Supplier-assortment-equipment	1	3	1	11	16
8. Equipment-assortment-supplier		6	4	6	16
9. Supplier-reception-order		6	1	9	16
10. Order-reception-supplier		3	8	4	15
11. Dept-issue-order		14	1	1	16
12. Order-issue-dept	2	4	6	4	16
13. Order-reference-equipment	1	2	5	7	15
14. Equipment-reference-order		6	3	6	15
15. Worker-hours_worked-timereg	1	4	5	6	16
16. Timereg-hours_worked-worker		2	6	8	16
17. Worktask-time_used-timereg	1	1	7	7	16
18. Timereg-hours_used-worktask		2	3	10	15
19. Project-project_time-timereg	1	1	6	7	15
20. Timereg-project_time-project	1	2	5	6	14
Total	**15**	**95**	**81**	**121**	

Table 5.4: Frequency table over the responses from the control group on experiment 1.

In table 5.5, which shows the frequencies from the treatment group, the most frequent choices are coloured grey. As can be seen, there are large differences for some tasks, and no differences for others. For instance, by comparing the responses from interpretation task 1 and 2 in the two tables, table 5.4 shows that the majority of respondents chose a connectivity of 0..N, while in table 5.5 the majority chose a connectivity of 1..N.

Interpretation tasks	Respons counts				Total
	0..1	0..N	1..1	1..N	
1. Supplier-agreement-dept	1	2	1	13	17
2. Dept-agreement-supplier		3		14	17
3. Dept-belong_to-employee		2	4	10	16
4. Employee-belong_to-dept	1	4	6	5	16
5. Dept-managed_by-manager	3	3	8	2	16
6. Manager-managed_by-dept	3	2	10	2	17
7. Supplier-assortment-equipment		8	2	7	17
8. Equipment-assortment-supplier	1	4	2	8	15
9. Supplier-reception-order		9		6	15
10. Order-reception-supplier	1	6	7	1	15
11. Dept-issue-order	1	11		4	16
12. Order-issue-dept	1	4	8	3	16
13. Order-reference-equipment	1	4	5	5	15
14. Equipment-reference-order	1	6	5	3	15
15. Worker-hours_worked-timereg		5	4	6	15
16. Timereg-hours_worked-worker		7	6	2	15
17. Worktask-time_used-timereg		4	2	7	13
18. Timereg-hours_used-worktask	1	3	4	5	13
19. Project-project_time-timereg		3	5	6	14
20. Timereg-project_time-project		4	3	7	14
Total	**15**	**95**	**82**	**116**	

Table 5.5: Frequency table over the responses from the treatment group on experiment 1.

Depicted graphically, the differences can be studied in figure 5.3, below:

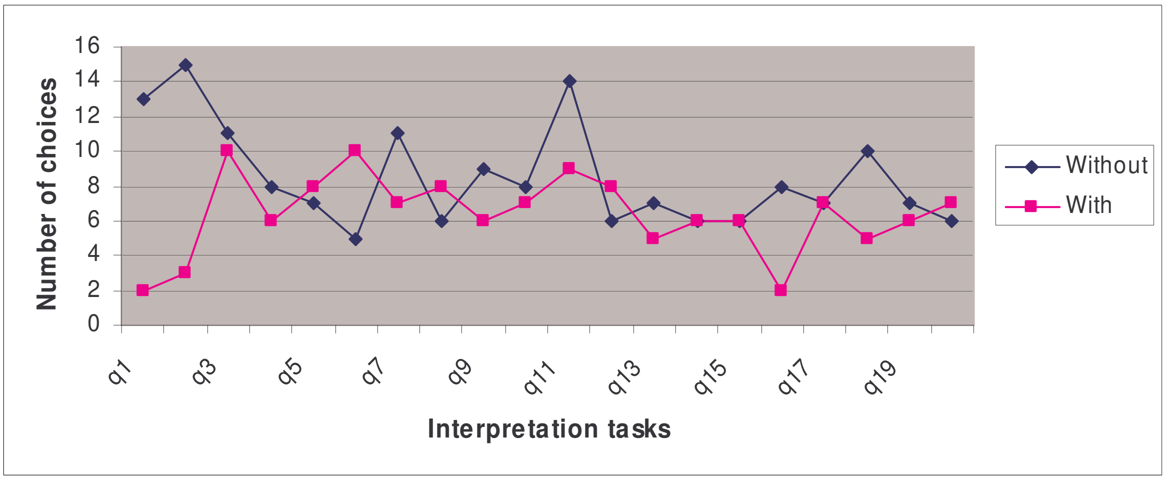


Figure. 5.3: Number of benchmark choices reported by the treatment group and the control group.

Figure 5.3 shows considerable variations with respect to the differences among the choices for each question. One reason for this may be that the membership conditions for some of the interpretation tasks were non-informative. Accordingly, the subjects in the treatment group would have to rely on the same kind of general knowledge that the control group used when making their judgements.

By separating interpretation tasks with informative membership conditions from those with non-informative membership conditions, it can be seen from figure 5.4 and figure 5.5 that for interpretation tasks with informative membership conditions most choices are opposite to each other, while for interpretation tasks with non-informative membership conditions, most choices are very close to each other so that the two curves exhibit almost the same pattern.

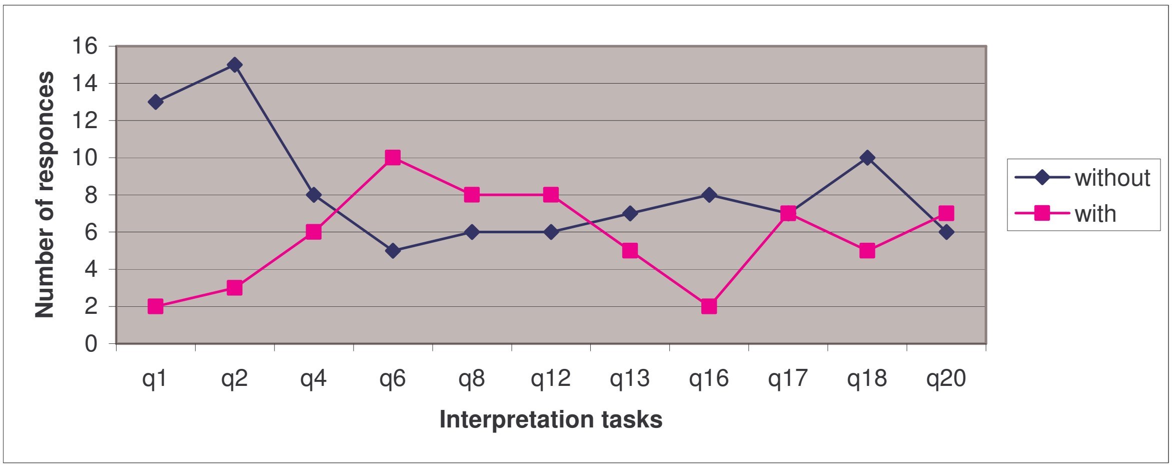


Figure 5.4: Number of benchmark choices for interpretation tasks with informative membership conditions.

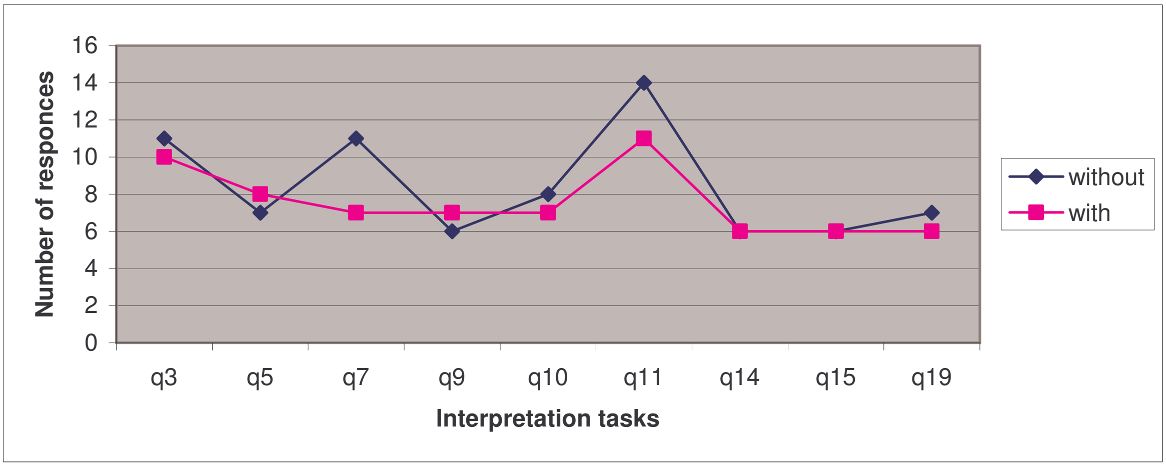


Figure 5.5: Number of benchmark choices for interpretation tasks with non-informative membership conditions.

Figure 5.6 depicts the mean confidence levels reported for each choice. Both curves show a declining tendency, where the subjects seem to be more confident in the start than in the end. This may be due to fatigue, or because the interpretation tasks become gradually more complex. The last twelve tasks are decomposed from two ternary relationships displayed in the full diagram, and may cause some confusion to less experienced designers.
However, it seems as if the subjects in the treatment group are generally more confident with the second half of the interpretation tasks, than are the subjects in the control group. The reason may be that the subjects in the treatment group, having no prior knowledge of membership conditions, needed some time to familiarize themselves with the new construct, in order to recognize the connection between membership conditions and confidence judgments. If this is the case, then it seems as if membership conditions may have a positive effect on the confidence in interpretations.

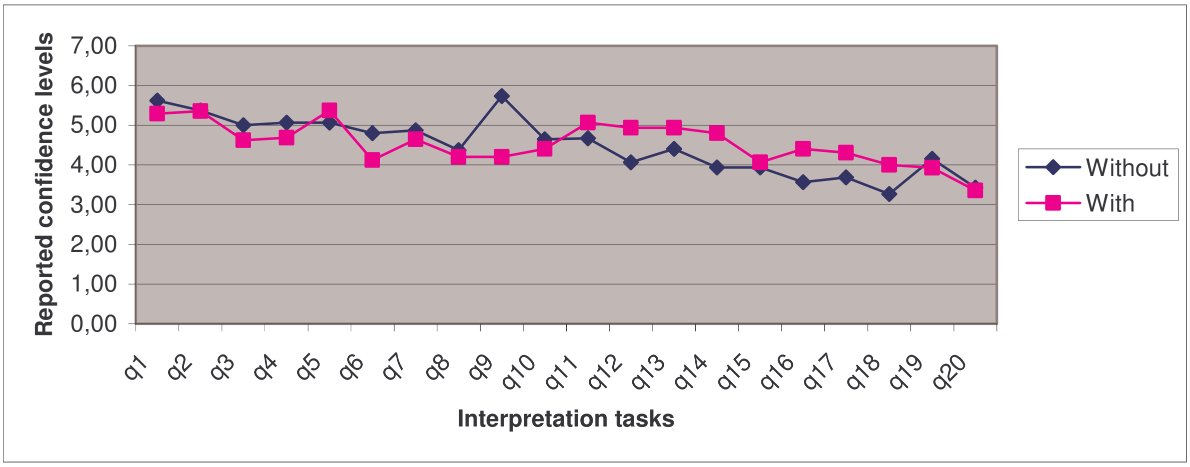


Figure 5.6: Reported confidence level for each choice.

Again, by separating interpretation tasks with informative membership conditions from those with non-informative membership conditions, a comparison of the second halves of figure 5.7 and figure 5.8 shows that the difference between the two groups is larger for the interpretation tasks with informative membership conditions than for those with non-informative membership conditions. In figure 5.8, in which both groups must base their interpretations on general knowledge, the two groups report almost similar confidence levels for most of the tasks.

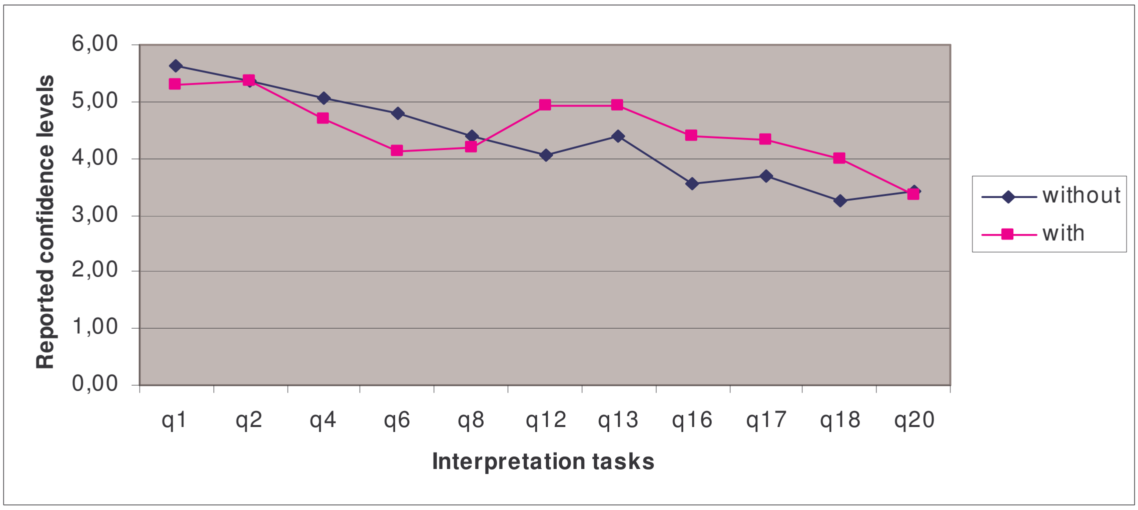


Figure 5.7: Reported confidence level for each choice based on informative membership conditions

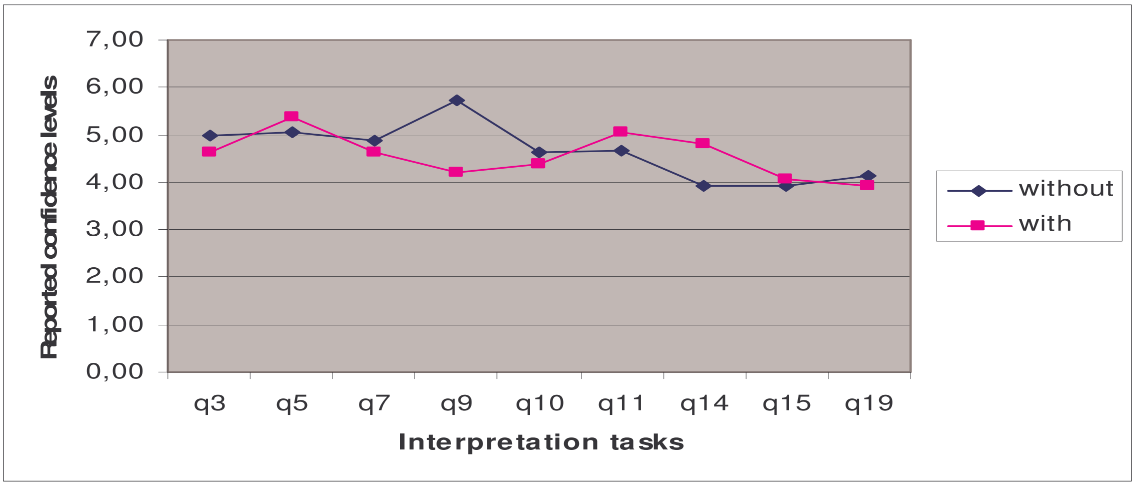


Figure 5.8: Reported confidence level for each choice based on non-informative membership conditions

Table 5.6 contains some descriptive statistics from experiment 1. The variables related to choice of connectivity are coloured grey in order to make the table easier to read. Based on the common sense benchmarks from the control group, the table shows differences between the two groups, both with respect to the choice of connectivity and the confidence in interpretations.

Considering the choice of connectivity first, there is an overall difference between the means in the two groups of 2.93. When only informative membership conditions are considered, the difference is 2.04, and for non-informative membership conditions the difference is 1.06. The variation in the responses, represented by the standard deviation, is larger for the treatment group on all three variables.

One reason for this may be that some of the subjects in the treatment group took greater notice of the membership conditions than others, who may have felt that reasoning based on general knowledge was more in line with the traditional approach taught in the course.

When it comes to the confidence, the differences between the means are smaller, and the variation is larger in the treatment group than in the control group. One explanation to this may be that the subjects in the treatment group were initially confused by the membership condition construct and did not learn to cope with it until the second half of the experiment. Another reason may be that some of the subjects in the treatment group took greater notice of the membership conditions, and thus felt more confident with their choices than others in the same group.

Statistics		**Treatment**	
		Without membership Conditions	**With membership Conditions**
Number of "correct" choices based on common sense benchmarks	Mean N Std. Deviation	10,69 16 2,15	7,76 17 3,17
Reported confidence for all tasks	Mean N Std. Deviation	4,49 16 1,19	4,40 17 1,47
Number of "correct" choices for tasks with informative membership conditions	Mean N Std. Deviation	5,69 16 1,49	3,65 17 1,90
Reported confidence for tasks with informative membership conditions	Mean N Std. Deviation	4,36 16 1,09	4,39 17 1,43
Number of "correct" choices for tasks with non-informative membership conditions	Mean N Std. Deviation	4,94 16 1,06	3,88 17 1,93
Reported confidence for tasks with non-informative membership conditions	Mean N Std. Deviation	4,63 16 1,37	4,38 17 1,56

Table 5.6: Descriptive statistics from experiment 1.

In order to see whether the differences in table 5.6 have any statistical significance independent sample t-tests were conducted. The results are depicted in table 5.7.

The t-tests were used to test the null-hypotheses presented in section 5.4.5.

Hypotheses		**Alternative hypothesis supported?**	**Significant Difference?**	***P*-value**
$H1_0$	Interpretation all tasks	Yes	Yes	0,004
$H2_0$	Confidence all tasks	No	No	0,838
$H3_0$	Interpretation informative membership conditions only	Yes	Yes	0,002
$H4_0$	Confidence informative membership conditions only.	No	No	0,948
$H5_0$	Interpretation non-informative membership conditions only.	Yes	No	0,064
$H6_0$	Confidence non-informative membership conditions only.	Yes	No	0,621

Table 5.7: Results from the t-tests on experiment 1.

The results from the t-tests show that, with α = 0.05, four of the alternative hypotheses were supported.

With a P-value of 0.004, hypothesis $H1_0$ seems highly unlikely. If the null-hypothesis were correct, a difference as large, or larger, than the one obtained in the experiment, can be expected to occur on average in only 1 time of 250. If we take into consideration that the subjects in the treatment group were inexperienced, and uninformed about membership conditions, the alternative hypothesis that membership conditions make a difference is clearly supported.

Hypothesis $H2_0$ is not supported by the t-test. That is to say that a P-value of 0.838 is not a sufficient reason to conclude that the two means differ. This result was unexpected, but on viewing the confidence reported for each question, there really was a difference. The trouble with it is that the treatment group is less confident than the control group during the first half of the tasks, and then shifts to be more confident during the second half. As already mentioned, the reason for this pattern, may be that the subjects in the treatment group needed some time to be acquainted with the membership condition construct, before they were able to use it properly. If this assumption is correct, a significant difference is expected to occur in experiment 2, where the subjects have learned about membership conditions in advance.

Hypothesis $H3_0$, with a P-value of 0.002 is even more unlikely than hypothesis $H1_0$. The fact that informative membership conditions make a larger difference between the means than non-informative membership conditions is a clear demonstration of the effect of membership conditions on interpretation tasks.

Concerning hypothesis $H4_0$, it was expected that there would be a difference between the means of confidence in the two groups. It was further expected that the difference would be larger for interpretation tasks with informative membership conditions than for non-informative membership conditions. However, since informative membership conditions were located in both the first and the second halves of the task set, the results became similar to the results in $H2_0$. Although the results did not support the alternative hypothesis to $H4_0$, it is expected that there will be a significant difference between the two means in experiment 2. In addition it is expected that the difference between the means based on informative membership conditions will be larger than the difference between the means for non-informative membership conditions.

With respect to hypothesis $H5_0$, it was expected that both groups would have to base their judgments on general knowledge, and that they would make similar choices. Based on the P-value from the t-test, it is not possible to conclude that there is any difference between the two means. This supports the alternative hypothesis that when only non-informative membership conditions are considered, the two groups will make similar choices. The reason is that both groups will have to rely on the same kind of general knowledge in order to solve the interpretation tasks.

Since it was expected that the two groups would use the same kind of general knowledge to guide their judgments, it was also expected that the two groups would report the same level of confidence in their choices. The alternative hypothesis to $H6_0$ is supported, since the P-value suggests that the null-hypothesis must be rejected.

Experiment 2

Experiment 2 was conducted the next day, after the subjects had received a lecture on concepts, classes, types, and membership conditions. The experiment followed the same approach as in experiment 1, but this time, the subjects that belonged to the treatment group in experiment 1 were placed in the control group in experiment 2, and those who belonged to the control group in experiment 1 were placed in the treatment group. The experiment gave the following results:

Interpretation tasks	**Respons counts**				**Total**
	0..1	**0..N**	**1..1**	**1..N**	
1.sup-shipment-delivery		7		7	14
2. delivery-shipment-sup		3	6	5	14
3. delivery-consist_of-subdel		9		5	14
4. subdel-consist_of_delivery		2	6	6	14
5. subdel-addressee-dept		5	7	2	14
6. dept-addressee-subdel		7	6	1	14
7. Subdel-content-medicine		8		6	14
8. medicine-content-subdel		5	1	8	14
9. med-medication-patient	1	8		5	14
10. patient-medication-med		8		6	14
11. patient-admission-dept	2	5	4	3	14
12. dept-admission-patient		8	1	5	14
13. patient-testing-allergytest	1	10		3	14
14. allergytest-testing-patient		7	2	5	14
15. patient-has-diagnose		11		3	14
16. diagnose-has-patient		6	3	5	14
17. disease-ref-diagnose	2	8	1	3	14
18. diagnose-ref-disease	1	5	1	7	14
19. doctor-states-diagnose		10		4	14
20. diagnose-states-doctor	1	4	3	6	14
21. employee-workplace-dept	1	3	2	8	14
22. dept-workplace-employee		3	4	7	14
23. doctor-manages-dept	3	8	2	1	14
24. dept-manages-doctor	3	5	5	1	14
Total	**15**	**155**	**54**	**112**	

Table 5.8: Frequency table of the control group responses in experiment2.

The most frequent choices made by the control group were selected as benchmarks and compared with the most frequent choices made by the treatment group. The differences can be seen by comparing the grey coloured cells in table 5.9 with the most frequent responses which are written in bold face in the same table. The differences are further depicted graphically in figure 5.11.

Interpretation tasks	**Responses**				
	0..1	**0..N**	**1..1**	**1..N**	**Total**
1. sup-shipment-delivery	1	1		12	14
2. delivery-shipment-sup	2	4	2	6	14
3. delivery-consist_of-subdel			2	12	14
4. subdel-consist_of-delivery		2	6	6	14
5. subdel-addressee-dept		1	8	5	14
6. dept-addressee-subdel	1	6	6	1	14
7. Subdel-content-medicine	2	1	4	7	14
8. medicine-content-subdel	1	6	1	6	14
9. med-medication-patient		8	1	5	14
10. patient-medication-med		10	3	1	14
11. patient-admission-dept	2	2	10		14
12. dept-admission-patient		11	2	1	14
13. patient-testing-allergytest	1	5	2	6	14
14. allergytest-testing-patient		3	3	8	14
15. patient-has-diagnose		7	3	4	14
16. diagnose-has-patient		7	4	3	14
17. disease-ref-diagnose	1	7	5	1	14
18. diagnose-ref-disease		6	4	4	14
19. doctor-states-diagnose		7		7	14
20. diagnose-states-doctor		3	2	9	14
21. employee-workplace-dept		1	8	5	14
22. dept-workplace-employee		1	4	9	14
23. doctor-manages-dept	5	3	3	3	14
24. dept-manages-doctor	2	4	5	3	14
Total	**18**	**106**	**88**	**124**	

Table 5.9: Frequency table of the treatment group responses in experiment2.

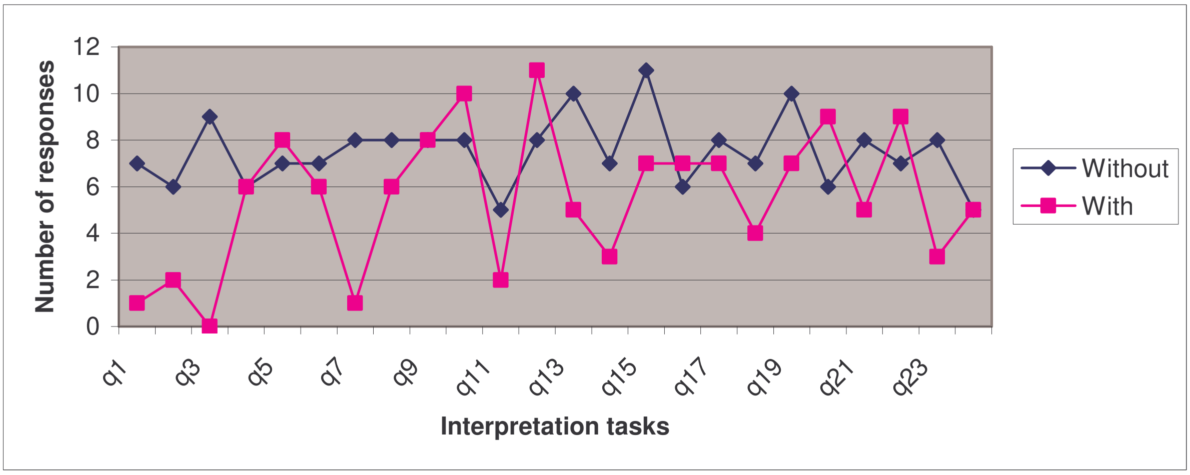


Figure 5.9: Number of benchmark choices reported by the treatment group and the control group.

Figure 5.10 and 5.11 depicts the choices made on interpretation tasks with informative and non-informative membership conditions respectively. Figure 5.10 exhibits the same kind of crisscrossing curve patterns as in figure 5.4 from experiment 1, showing considerable differences between the two groups. In figure 5.11, on the other hand, the differences seems to be somewhat larger compared to those in figure 5.5 from experiment 1. Still there are no differences as large as the largest ones in Figure 5.10.

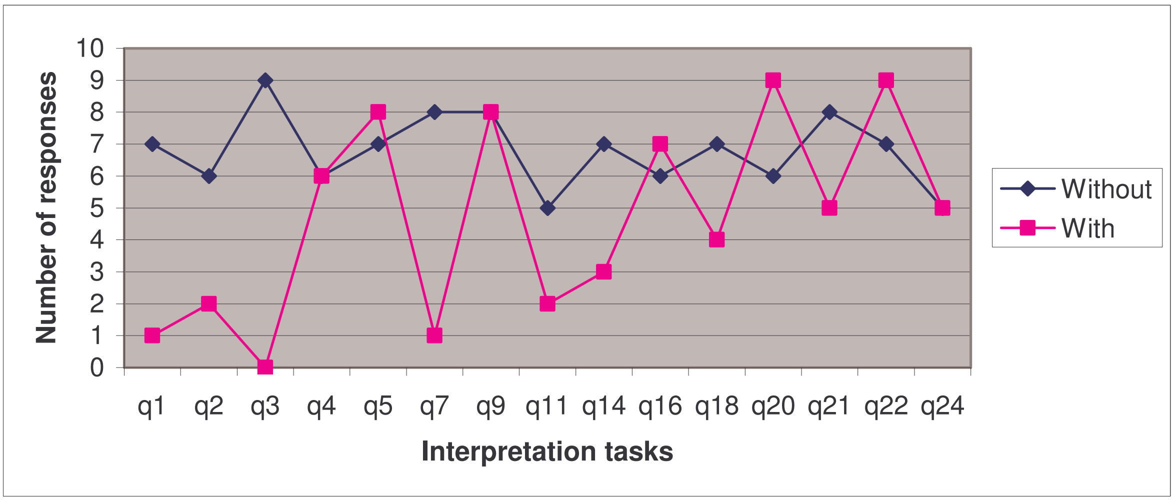


Figure 5.10: Number of benchmark choices for interpretation tasks with informative membership conditions.

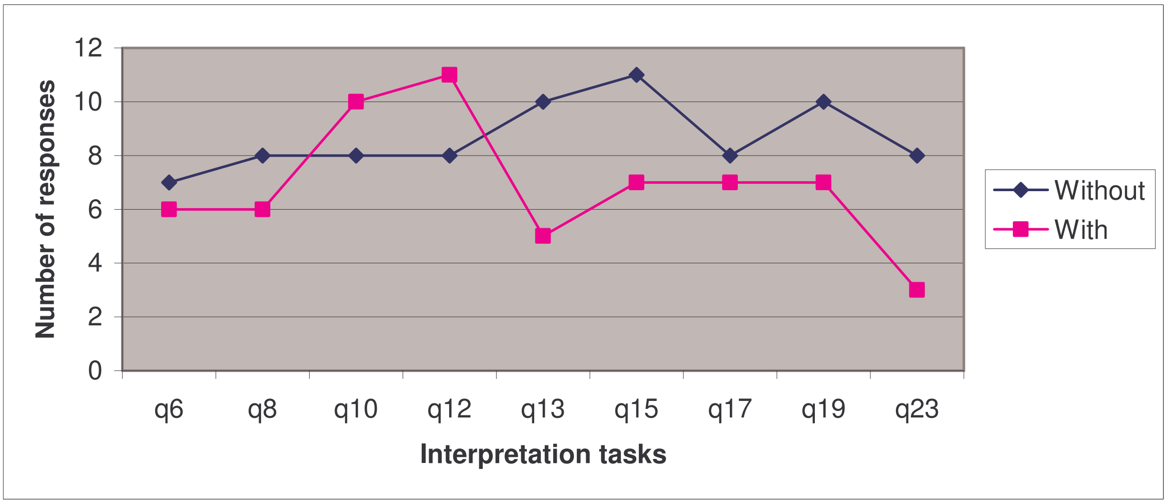


Figure 5.11: Number of benchmark choices for interpretation tasks with non-informative membership conditions.

Figures 5.12, 5.13 and 5.14 depict the mean confidence levels. As can be seen, the treatment group is more confident in all choices. The main reason for this seems to be that the control group is consistently less confident in their choices in experiment 2 than in experiment 1.

Without knowing the membership conditions, the subjects in the control group have become more careful, or conservative in their judgments.

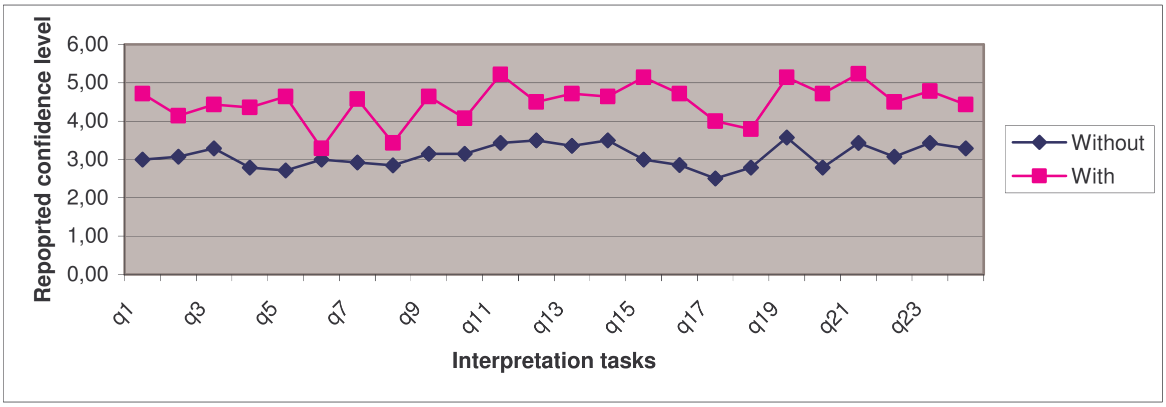


Figure 5.12: Reported confidence level for all choices.

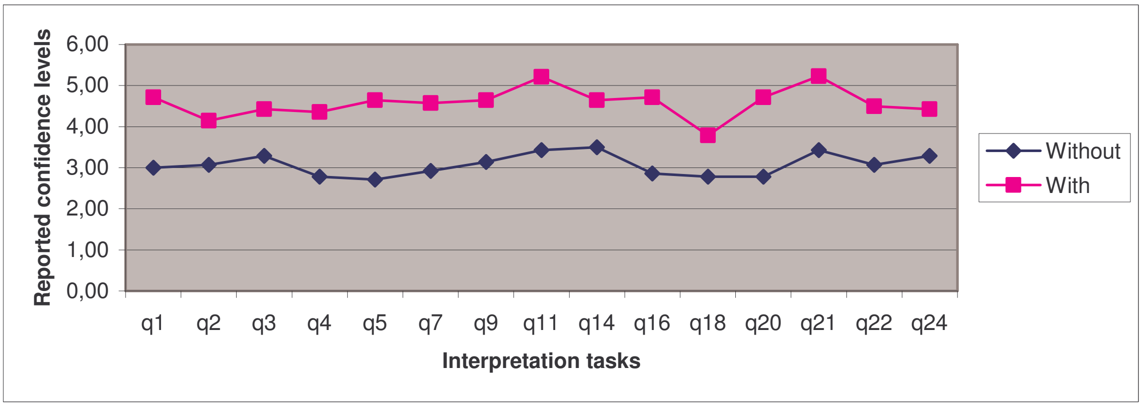


Figure 5.13: Reported confidence level for each choice based on informative membership conditions.

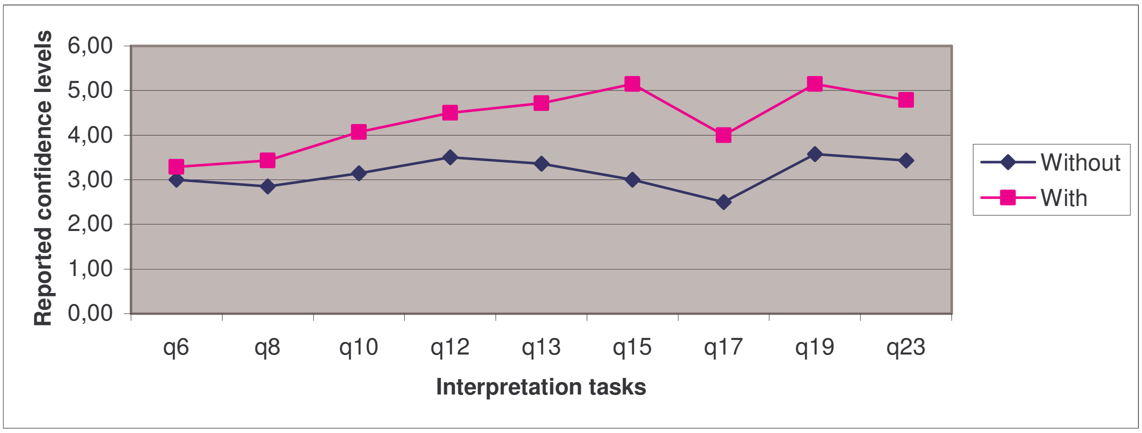


Figure 5.14: Reported confidence level for each choice based on non-informative membership conditions.

Table 5.10 contains some descriptive statistics from experiment 2. Compared with the statistics in table 5.6, of experiment 1, a number of differences can be noted: With respect to the choice of connectivity, the difference between the means in the two groups has increased for all choices taken together, as well as for choices based on informative membership conditions. For choices based on non-informative membership conditions, the difference has become smaller. In addition, for all choices taken together, and for choices based on informative membership conditions, the variation in the responses of the control group has increased, while it has decreased in the treatment group. For choices based on non-informative membership conditions, the variation has increased in the control group, while it has remained the same in the treatment group.

Also, when considering the confidence in interpretations, the numbers show a major decrease in the means for the control group, along with an increase in the variation.

Taken together, the results seem to indicate that membership conditions will influence the interpretation of relationships in two ways: If the interpretation of data models can be based on membership conditions, the results will be different from interpretations based on general knowledge only. In addition, membership conditions seem to result in more consistent judgments between subjects, in the sense that the variation becomes smaller. This is especially evident from the results involving informative membership conditions. Finally, when interpretations are based on general knowledge the subjects are less confident and less consistent in their interpretations.

Statistics		**Treatment**	
		Without membership Conditions	**With membership Conditions**
Number of correct choices based on common sense benchmarks	Mean	12,86	9,43
	N	14	14
	Std. Deviation	2,69	2,95
Reported confidence for all tasks	Mean	3,09	4,49
	N	14	14
	Std. Deviation	1,70	1,54
Number of correct choices for tasks with informative membership conditions	Mean	7,43	4,79
	N	14	14
	Std. Deviation	2,14	1,80
Reported confidence for tasks with informative membership conditions	Mean	3,08	4,58
	N	14	14
	Std. Deviation	1,68	1,58
Number of correct choices for tasks with non-informative membership conditions	Mean	5,36	4,64
	N	14	14
	Std. Deviation	1,69	1,95
Reported confidence for tasks with non-informative membership conditions	Mean	3,15	4,36
	N	14	14
	Std. Deviation	1,77	1,54

Table 5.10: Descriptive statistics from experiment 2.

As for experiment 1, independent sample t-tests were used to test the null hypotheses. The results from the t-tests are depicted in table 5.11.

Hypotheses		**Alternative hypothesis supported?**	**Significant Difference?**	***P*-value**
$H1_0$	Interpretation all tasks	Yes	Yes	0,003
$H2_0$	Confidence all tasks	Yes	Yes	0,031
$H3_0$	Interpretation informative membership conditions only	Yes	Yes	0,002
$H4_0$	Confidence informative membership conditions only.	Yes	Yes	0,022
$H5_0$	Interpretation non-informative membership conditions only.	Yes	No	0,310
$H6_0$	Confidence non-informative membership conditions only.	Yes	No	0,064

Table 5.11: Results from the t-tests on experiment 2.

From table 5.11, it can be seen that with $\alpha = 0.05$, all six alternative hypotheses are supported. There is a significant difference between the means of all interpretation tasks obtained for the treatment group and the control group, so hypothesis $H1_0$ can be rejected.

Similarly, there is a significant difference with respect to the confidence reported by the two groups, so hypothesis $H2_0$ can be rejected as well.

As expected, hypotheses $H3_0$ and $H4_0$ can also be rejected. Being concerned only with informative membership conditions, the differences between the means are more significant than for $H1_0$ and $H2_0$.

With respect to $H5_0$ and $H6_0$, being concerned with non-informative membership conditions, it was expected that the difference would be insignificant, since both groups would have to rely on the same kind of general knowledge when interpreting the relationships. As indicated by the results, there is no reason to conclude otherwise, so both $H5_0$ and $H6_0$ can also be rejected.

5.5 Overall interpretation of the results

The research question for this study was to measure the effect of membership conditions on the interpretation of relationships in a conceptual model. It was expected that knowledge of membership conditions would have an effect on the choice of interpretation, as well as on the confidence in the interpretation. The results from the study show that membership conditions seem to affect the interpretation of relationships in several ways. First, it is shown that membership conditions may influence the interpretation of relationships, in the sense that people who know the membership conditions will make different judgments than people for which the membership conditions are unknown.

Second, it is also shown that without knowledge of membership conditions people become less confident, and also less consistent in their interpretations, exhibiting larger variation in their judgments.

The practical significance of this is that in a real situation, knowledge of membership conditions may enhance the reading, understanding and validation of conceptual models, and make the readers more confident that their interpretations correspond with the designers' intentions.

5.6. Conclusion and future research

In this study the basic assumption that membership conditions have an important function in the interpretation and validation of conceptual models is supported by six different, though related hypotheses. It should be noted though, that with such low sample sizes as n=33 for experiment 1, and n=28 for experiment 2, care should be taken not to draw any premature generalizations from the findings. This suggests that the experiment should be repeated with larger samples, different tasks, and with expert designers as well as with novices. For a more detailed discussion of the study's validity and reliability, see section 6.3.3 on page 148.

With respect to interpretation, knowledge of membership conditions may help IT-auditors, users and designers to better understand and validate conceptual models made by others.

6.0 Overall discussion

6.1 Introduction.

This thesis consists of three studies that are all based on the assumption that classification has not received sufficient attention by the data modelling community. The first study, which is described in chapter 2, was conducted to analyze the concept of classification. The study uncovered four different senses of classification, and lead to the analyses of related concepts, such as *concept*, *class*, *type*, *object* and *property*.

Having found an answer to what classification may mean to conceptual modelling, the second study, which is described in chapter 3, was conducted to test the initial assumption that classification is not properly attended to by the discipline. Based on a content analysis of 29 text books on conceptual modelling and database design, it was shown that none of the text books contained explicit definitions of classification in all four senses.

Based on the findings from the first two studies, a methodology that integrates classification and conceptual modelling was developed and presented in chapter four. The chapter provides a theoretical justification for a constructivist perspective on classification and conceptual modelling, explains its theoretical concepts, and describes the method through a set of guidelines and examples. The method is also a demonstration of the pragmatic utility of the conceptual framework developed in the first study.

One of the implications from the first study is that classification may affect the interpretation and the design of conceptual models. The third study, which is described in chapter 5, was conducted to empirically test the effect of classification on *interpretation* tasks. It is shown that people who know the membership conditions will make other judgments than people for which the membership conditions are unknown. It is also shown that without knowledge of membership conditions people become less confident, and less consistent in their interpretations, exhibiting larger variation in their judgments.

6.2 Implications for theory and practice.

Chapter 2 identified the following implication of classification:

1. Shared understanding of basic concepts;
2. Formal verification of completeness, logical consistency and understandability;
3. Basis for modelling decisions.
4. Socio-technical consequences and measures;
5. Data integrity;
6. Validation and interpretation of conceptual models;
7. Schema integration

The implications were first described in section 2.5, starting on page 56. What was said there will not be repeated. Instead, supplements will be made, based on insights gained from studying the implications as reported in chapter 3 through 5. Suggestions for further research will also be added.

6.2.1 Shared understanding of basic concepts.

This is basically an implication for theory. The findings in chapter 2 and 3 suggest a close interdependency between concepts such as classification, concept, class, type, object and property. The concepts contribute to each other's meaning, and constitute a coherent concept system that may be used to negotiate and advance a unified vocabulary for conceptual modelling.

The ease with which the concepts can be presented and used is demonstrated in chapter 4. The extent, to which students, practitioners, and researchers experience the concept system as meaningful and useful, remains to be explored.

6.2.2 Complete, logical consistent and understandable vocabulary.

The classification process results in a vocabulary. The theoretical implication of classification in this respect is that completeness and logical consistency can be formally defined. The practical implementation is that tools can be developed to perform automatic completeness and consistency checks. Research on so called taxonomic reasoning has been reported by Bergamaschi and Sartory (1992), and may suggest directions for further research on this topic.

6.2.3 Basis for modelling decisions.

This is an important implication for practice. The vocabulary, its logical and hierarchical structures may be used to guide the naming, selection and justification of constructs such as value-sets, entity types, relationship types, attributes and roles in the conceptual model.

It is not only the guiding that is important, but even more so, the possibility of unearthing modelling decisions. By insisting that deviations from the vocabulary must be explicitly justified, design decisions are made explicit. Hence, a conceptual model may be supplemented with an explanation for why it looks the way it does. This aspect is demonstrated in chapter 4, and can be further explored by case studies and experiments on design tasks. Appendix D contains an instrument to study the effects of classification on design tasks.

6.2.4 Socio-technical consequences and measures.

Classification draws attention to the socio-technical *consequences* of misclassification, and to the *organizational*, *administrative*, and/or *technical* measures that need to be taken to avoid unwanted consequences.

From a practical point of view, classification makes a strong connection between the micro process concerning the database design and the macro process that covers the complete systems analysis and design process. Classification is not only informed by the macro process. It actually informs the macro process about socio-technical aspects that are uncovered during the classification process.

6.2.5 Data integrity

Membership conditions may, in principle, be associated with each concept in the conceptual model. The theoretical implication of this is that membership conditions represent a general enhancement to data integrity by offering an opportunity to specify a new kind of integrity constraint associated with types.

The practical implication is that the membership conditions need to be formalized and transformed from linguistic expressions in the vocabulary to user instructions and/or equivalent expressions in data and/or programming languages. One example of this transformation stream is briefly sketched in chapter 4, but further research and development is needed with respect to language extensions and transitions between conceptual, logical and physical languages.

6.2.6 Validation and interpretation of conceptual models.

Classification adds to the semantics represented by a conceptual data model. Knowledge of the membership conditions may help auditors, users, and systems analysts to validate and/or interpret existing models. The effect of membership conditions on interpretation tasks is demonstrated in chapter 5. Further research should be conducted by replicating the experiment in new settings.

The theoretical implication of this is that knowledge of membership conditions may enhance the reading, understanding and validation of conceptual models, and make the readers more confident that their interpretations correspond with the designers' intentions.

The practical implication is to consider how the membership conditions can best be represented in the conceptual model.

6.2.7 Schema integration.

Schema integration is the process of generating one or more integrated schemas from existing schemas. The process can be broken down into several phases, such as the framework described by Ram and Ramesh (1999), page 122. Since membership conditions represent an addition to the semantics normally considered part of a database schema, it may be of both theoretical and practical interests to consider if, and how such knowledge may contribute to solve the problems associated with each phase. Although recent surveys on schema matching approaches mention linguistic approaches, (Rahm, 2001), membership conditions are not considered.

6.3 Validity and reliability

Validity and reliability are two central concepts in assessing the quality and rigour of quantitative research. According to Sarantakos, (1998), *validity* refers to the ability of an instrument to measure what is supposed to be measured, while *reliability* refers to the ability of an instrument to produce consistent results. In qualitative research, alternative concepts have been suggested, such as *credibility*, *transferability*, *dependability* and *confirmability*, (Polit and Hungler, 1999). Table 6.1 shows how the various concepts are related and what they mean.

Criteria for assessing quantitative measures	Criteria for assessing qualitative measures	Description
Internal validity	Credibility	Internal validity, or credibility, is concerned with the credibility of the data and the conclusions. Do the instruments and findings make sense?
External validity	Transferability	External validity, or transferability, refers to the generalizability of the data. Are the findings applicable to other contexts?
Internal reliability	Dependability	Internal reliability, or dependability, refers to the stability of data over time and across researchers and methods. Can the study be replicated?
External reliability	Confirmability	External reliability, or confirmability, refers to the objectivity or neutrality of the data. Do two or more observers agree on the data's relevance or meaning?

Table 6.1: Criteria for assessing the quality and rigour of research.

Since the first two studies, reported in chapter 2 and 3 are interpretative in nature, their quality and rigour will be discussed according to the qualitative measures described in table 6.1.

6.3.1 Concept analysis of classification.

The first study was conducted to explore the meaning of 'classification'. To enhance the *credibility*, the study was based on *triangulation*. The basic idea behind triangulation is that the more agreement among different data sources the more reliable is the interpretation of the data. Data were collected from multiple sources, (individuals, disciplines, and organizations), from different points in time, (ranging from 1971 till 2002), from multiple sites, (journals, proceedings and text books), and by different methods, (key word searches, author lists).

This can be viewed as multiple triangulation, though some would probably argue that to call this approach triangulation is an overstatement, since all papers were collected via a single application and in a single process, and that papers represent a single kind of information source, not multiple sources. They are probably right. To really deserve to use the term triangulation, the data collection could have been extended with interview data for instance. Still, the data that actually were collected reflect a diversity that should make the study credible.

When it comes to the *transferability* of the findings, the conceptual framework developed in chapter 2, as well as the method described in chapter 4 are deemed sufficient to enable readers to evaluate the applicability to their own settings and groups.

No measures have been taken to assess the *dependability* of the study, but steps have been taken to describe, in as much detail as possible, how the search process was carried out. In spite of this, it is doubtful if a new study would produce exactly the same data. First, the precision and recall may be affected by new literature being added to the databases, or by changes being made to the user interface and search algorithms of the database application. Second, the actual selection of documents depended on my subjective interpretation of titles, abstracts and key words, there and then. Other researchers would probably make other choices according to their personal knowledge and perspectives. Third, during the analysis, which lasted for more than a year, the annotations that were taken, and, later, the interpretations of the annotations, were constantly affected by what I learned about the subject matter.

As for the *confirmability*, the search process started off with a few, broad terms such as “classification”, “concept”, and “class”. The search terms were then gradually specialized based on the keywords that characterized the retrieved papers. This was done to protect the search process from being influenced by any preconceived or premature ideas of classification. In addition to keyword searches, a search process was performed based on references to persons such as keynote speakers at conferences, and authors of invited papers. To further discern the documents from any possible researcher bias, a sample of 115 out of a total of 288 papers were randomly selected for the final analysis. Lastly the sample was increased with 12 papers considered to be classic, specially invited, or surveys.

Lastly, chapter 4 demonstrates the *pragmatic utility* of the conceptual framework.

Based on the purpose stated in section 2.1, the following objectives were specified:

1. Clearly reflect the meaning of classification as it pertains to conceptual modeling.
2. Provide guidelines on how to use classification in conceptual modeling.
3. Provide guidelines on how to evaluate the results of classification.
4. Contribute to the development of a coherent vocabulary for classification and conceptual modeling.

The meaning of classification is presented in section 4.3, while guidelines on classification and evaluation are given in section 4.4. In the first run, it must be up to the readers to decide whether the objectives have been reached or not, but the intention is to have the objectives empirically tested.

6.3.2 Content analysis of text books on conceptual modelling.

The second study was conducted to test the assumption that classification was not properly attended to by the conceptual modelling tradition. Based on concept definitions arrived at in the first study, eleven different, but related concepts, which all had to do with classification, were coded, sought for, and tallied, in a content analysis of 29 text books.

To strengthen the *credibility*, the search process, as well as the analysis, was based on a principle called *search for disconfirming evidence*. Instead of searching for text books that might support the initial assumption, the search process was based on terms from table 2.2 and 2.3 in combination with terms like “Practical approach?”, “Introduction”, “Advances”, and so on. In this way, text books would be retrieved if they contained terms in their title, abstract or keyword lists that indicated a certain awareness of classification. In addition, the analysis was based on very liberal interpretations of the texts, in order not to confirm the initial assumptions on insufficient grounds.

What about the *transferability*? The findings are based on text books written by writers who are considered to be well established in the field of conceptual modelling. If this assumption is correct, then it should be possible to generalize from the sample to the total population, which, in this context, means the field of conceptual modelling in general.

Can the study be *replicated*? The answer is both yes and no. The books are readily available for a second analysis, and the terms to look for are well described and exemplified. Still, there is an interpretative factor that requires a certain understanding of classification in order to recognize the terms or variables when they are only implicitly represented in the texts.

To enhance the dependability, the study could have been designed with a larger emphasize on quantitative measures. Instead of interpreting the text in each book, the glossaries and/or indexes could be searched in ten times as many books. The results would certainly be more dependable, but at the same time more meagre, compared to the thick descriptions that result from the interpretative approach.

With respect to the *confirmability* of the data, the text books have been selected by using a list of writers that I consider to be well recognized in the field. They have held various key positions at conferences, edited proceedings, reviewed papers, authored and co-authored text books, been referred to in many surveys and bibliographies, and hold teaching and/or research positions at universities and research centres. However, since the list is compiled by me, it might just be the case that most of these writers have a special thing in common, and that is not to focus too much on classification in their books. There may be other writers, who enjoy just as much recognition in the field as "my writers" do, and who focus expressively on classification. In order for the reader to assess the confirmability of the data, the list of authors is made explicit, and the findings have been demonstrated by extensive citations from the text books.

6.3.3 Effects of classification on interpretation tasks.

Chapter five describes an experimental study to empirically test the effect of classification on interpretation tasks. It was found that knowledge of membership conditions had an effect on the subjects' judgments of relationships in ER-diagrams, as well as on the confidence associated with their judgments. However,

The experiment is *internally valid* if it can be demonstrated that the observed effect is not caused by conditions other than knowledge of membership conditions. One condition which cannot be ruled out is researcher bias. The experiment was designed, administered, recorded, and analyzed by the same person, the researcher. Accordingly, the researcher's expectations may have influenced the ways the data are interpreted. There may be patterns in the data that have been overlooked, or there may be patterns that have been over emphasized.

In retrospect, the notion of "knowledge of membership conditions" should have been explicitly defined. During the first experiment "knowledge of membership conditions" was taken to mean *access* to membership conditions associated with the interpretation tasks. During the second experiment, a second component was added to the concept, so that "knowledge of membership conditions" now was understood as 1) *access* to membership conditions and 2) *increased awareness* of the membership conditions. Since all participants in the second experiment were given the same lecture, the subjects' characteristics were changed between the two experiments, making the results from the two experiments less comparable. A less confusing design would be to run a single experiment where the treatment group receives a lecture on membership conditions as well as access to membership conditions, while the control group receives a stripped version of the interpretation tasks only.

With respect to the *external validity* of the experiment, the small sample is not representative for the total population. The subjects represent a convenience sample taken from a single class of undergraduate students, at a single department, at one university. Further more, the design, administration, data recordings, and analysis have been carried out by a single person, the researcher. To strengthen the external validity of the experiment, it should be repeated, with larger samples taken from diverse institutions, more researchers, and more tasks. Hence, claims based on generalizations from the experiment, would be premature and in the worst case, fallible.

Concerning the *internal reliability,* no measures have been taken to assess the instrument's stability and internal consistency, because of the small sample size. However, the experiment should be easy to replicate. The process is fully described, the instrument is included in appendix A, the lecture, which makes up a part of the independent variable, is included in appendix B, and the SPSS output files in appendix C.

As an alternative design, it may be considered to run a single experiment where the treatment consists of a lecture on membership conditions as well as access to membership conditions. The treatment is given to the treatment group only. This design is simpler and less confusing with respect to what exactly constitute the independent variables.

Lastly, the *external reliability* of the experiment can be discussed. Some may argue that the instrument used is better suited to measure design tasks, in that the subjects are presented to incomplete models that they are asked to complete. On the other hand, the external reliability is supported by the fact that other research projects, reported by Siau, Wand and Benbasat (1995), (1997), and by Dunn and Grabski (2001), make use of similar instruments to measure interpretation tasks.

It is possible to redesign the instrument, so that each task is represented by a separate and complete model, and to have the respondents answer if the model is correct or incorrect. This would clearly strengthen the face validity of the instrument, and possibly also eliminate any disagreements among researchers about the instrument's significance to the study situation. However, to cover the twenty tasks represented in the current instrument, four times as many tasks would be required in the modified version. This may cause an extra cognitive burden on the respondents, which, in turn, may threaten the instrument's internal validity.

6.4 Recommendations and future research

The findings reported in this thesis suggest that classification plays a central role with respect to conceptual modelling. Still, more research is needed in order to corroborate or dismiss the findings. In addition to the research questions associated with each implication in section 6.2, the most immediate recommendation is to continue work on three ongoing projects:

1. The first project is to offer students to replicate the study of interpretation tasks in different settings and possibly by different research designs. In addition to use an experimental design, case studies may be used to collect supplementary data on how membership conditions affect the reasoning process during interpretation tasks.

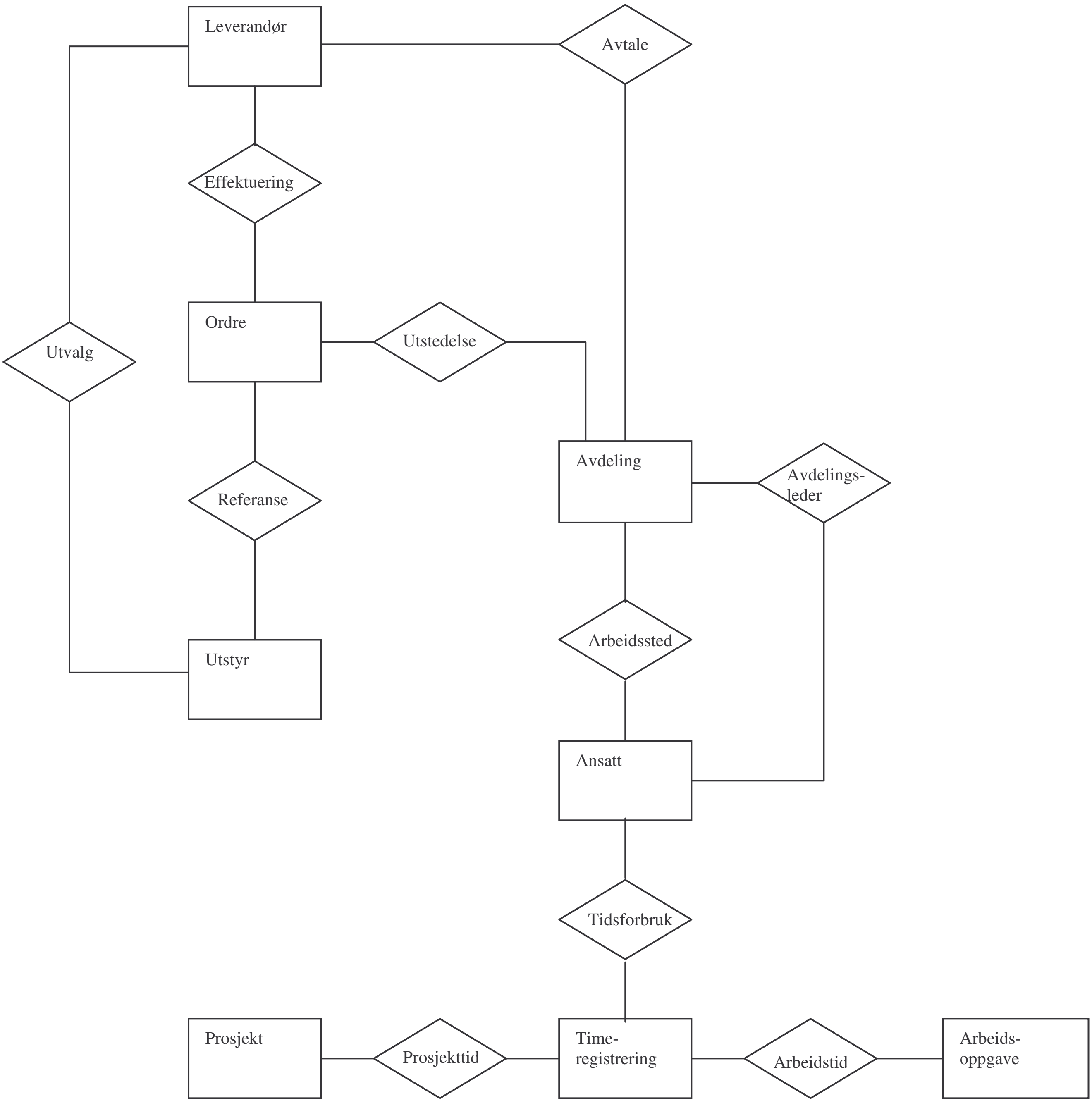
2. The second project is to study the effect of classification on design tasks. The control group will receive a requirement specification only, while the treatment group will receive the requirement specification and a vocabulary.
   Appendix D contains instruments based on design tasks created and used by Shoval and Shiran (1997). The general idea is to measure 1) the correctness of the conceptual schemas being designed, 2) time to complete the designs, and 3) designers' confidence in their designs. This makes it possible to compare the results with Shoval and Shirans measures.

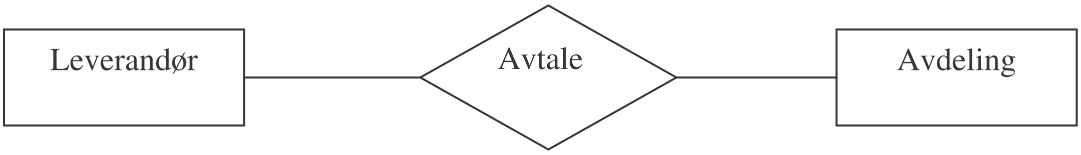
3. The third project is to develop a database application to store metadata about existing application terminologies. The meta database is supposed to guide the selection of terms during design of new applications, and to provide different perspectives on term usage during integration of applications. Appendix E contains a paper that reflects some initial ideas about the topic.

Appendix A

Instrument to measure interpretation tasks

Experiment 1





1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Avtale.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Avtale.

A. 0,1

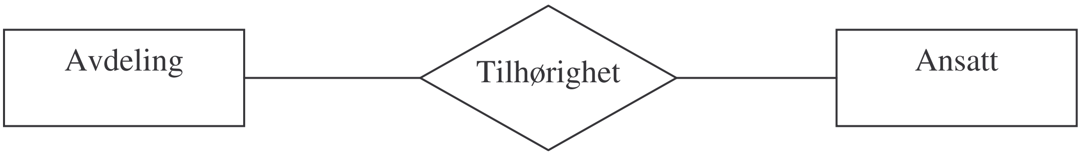
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Tilhørighet.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ansatt i forhold til relasjonen Tilhørighet.

A. 0,1

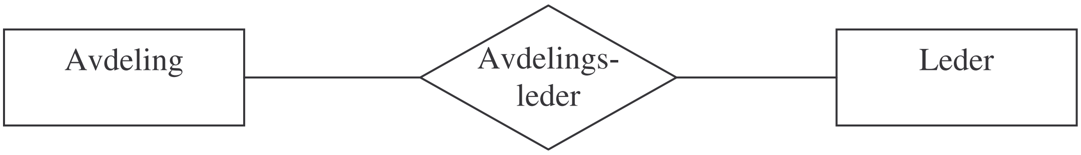
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Avdelingsleder.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdelingsleder i forhold til relasjonen Avdelingsleder.

A. 0,1

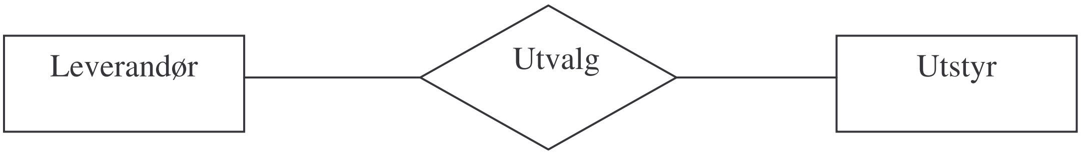
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Utvalg.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Utstyr i forhold til relasjonen Utvalg.

A. 0,1

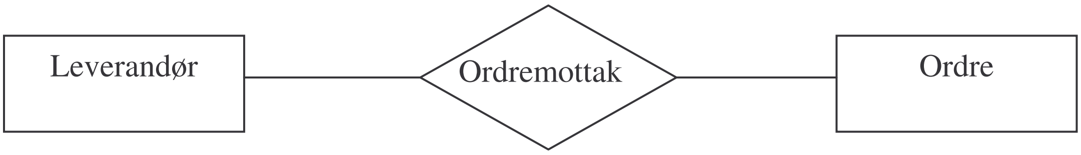
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Ordremottak.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ordre i forhold til relasjonen Ordremottak.

E. 0,1

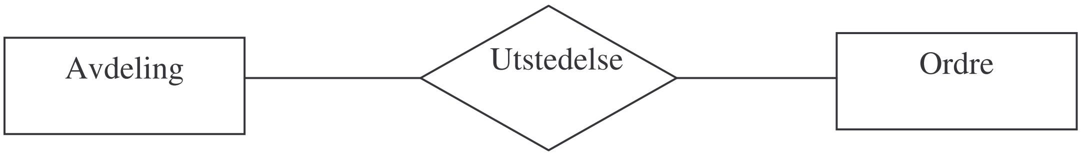
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Utstedelse.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ordre i forhold til relasjonen Utstedelse.

A. 0,1

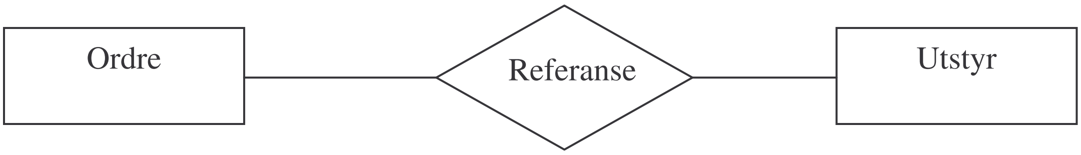
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ordre i forhold til relasjonen Referanse.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Utstyr i forhold til relasjonen Referanse.

A. 0,1

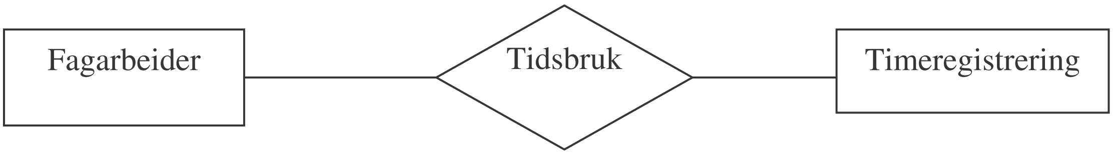
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Fagarbeider i forhold til relasjonen Tidsbruk.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Timeregistrering i forhold til relasjonen Tidsbruk.

A. 0,1

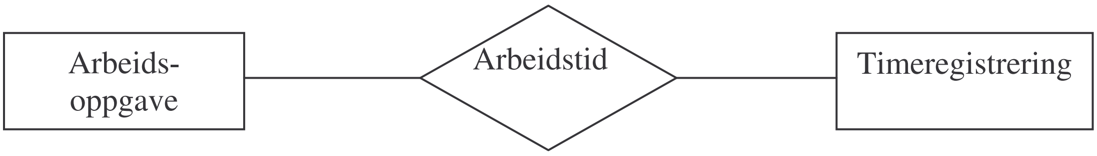
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Arbeidsoppgave i forhold til relasjonen Arbeidstid.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Timeregistrering i forhold til relasjonen Arbeidstid.

A. 0,1

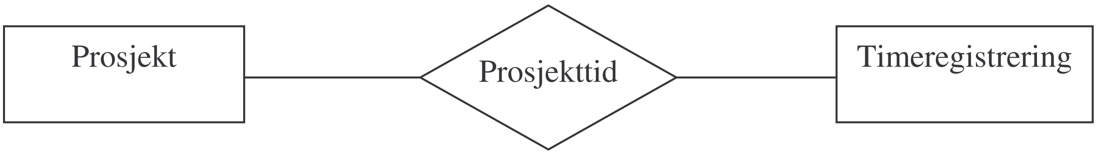
B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Prosjekt i forhold til relasjonen Prosjekttid.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Timeregistrering i forhold til relasjonen Prosjekttid.

A. 0,1

B. 0,N

C. 1,1

D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

Spørreskjema for demografiske data

1. Hvor erfaren er du med datamodellering?

Ikke erfaren	1	2	3	4	5	Erfaren

2. I hvilken grad er du fortrolig med ER-modellen?

Ikke fortrolig	1	2	3	4	5	Fortrolig

3. I hvilken grad har du erfaring med ER-modellering?

Ingen erfaring	1	2	3	4	5	Mye erfaring

4. I hvilken grad er du fortrolig med den syntaksen som er benyttet i testen?

Ikke fortrolig	1	2	3	4	5	Fortrolig

5. Hvor mange semester har du studert ved UiB? ____ semester inkludert inneværende semester.

6. Dersom du har relevant praksis i forhold til studiet, oppgi antall praksisår: ____ år.

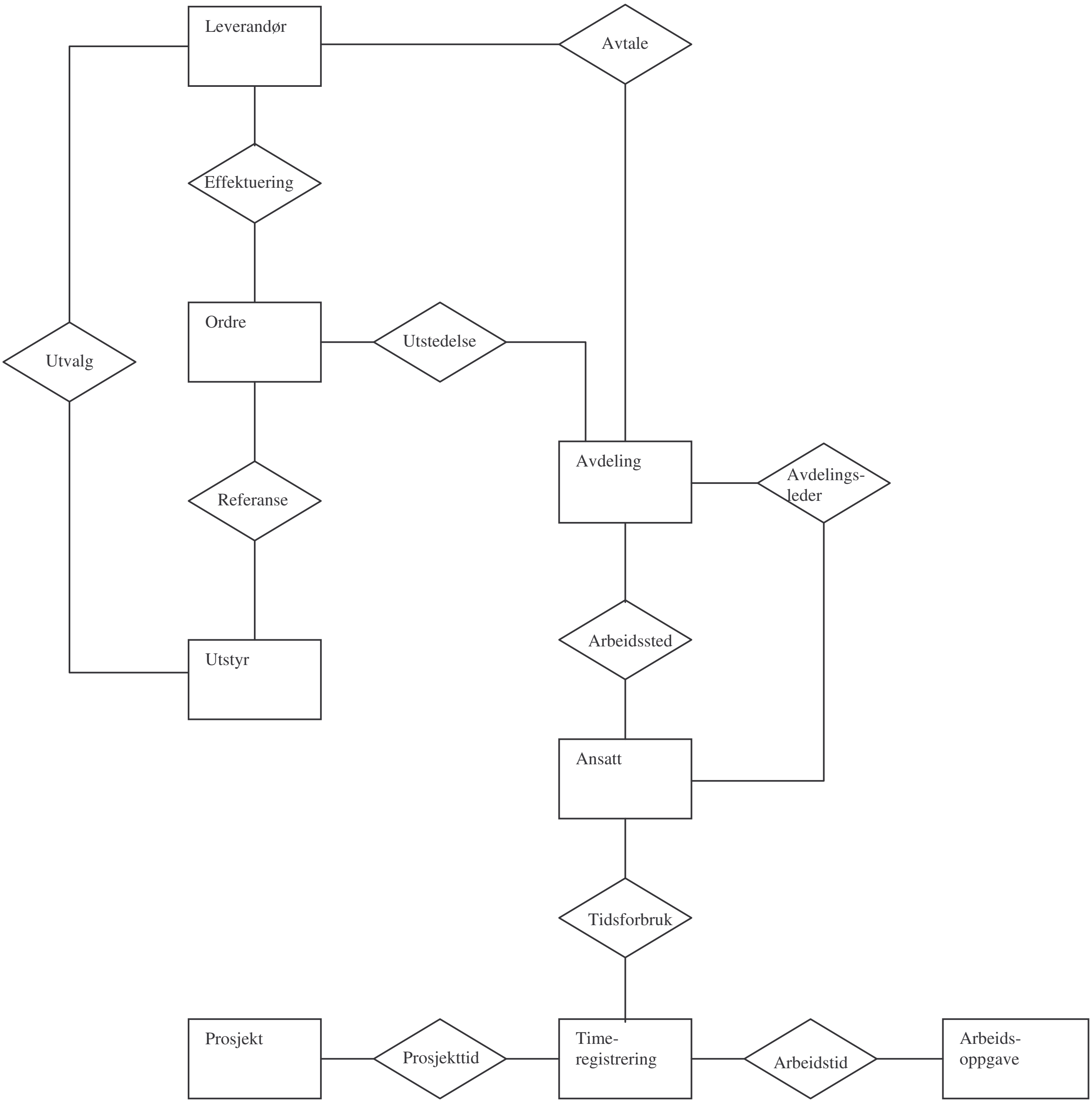
7. Hvor gammel er du? ____ år.

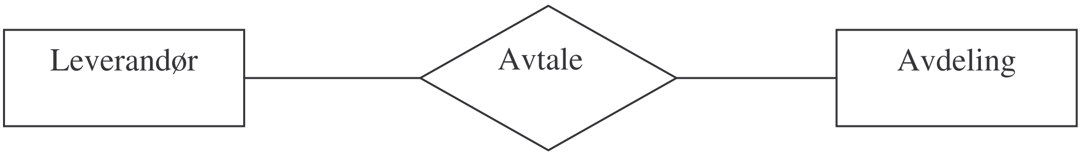
8. Kjønn

Mann □

Kvinne □

Experiment 2





En leverandør er et firma som har inngått en leveringsavtale med minst en avdeling.

En avdeling er en organisasjonsenhet som har en leveringsavtale med minst en leverandør.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Avtale.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Avtale.

E. 0,1

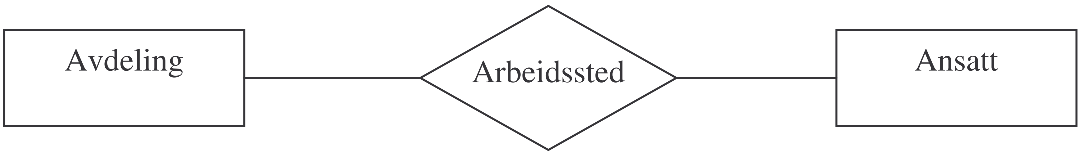
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En avdeling er en organisasjonsenhet som har en leveringsavtale med minst en leverandør.

En ansatt er en person som jobber som leder, ingeniør, fagarbeider eller sekretær i en avdeling.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Arbeidssted.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ansatt i forhold til relasjonen Arbeidssted.

E. 0,1

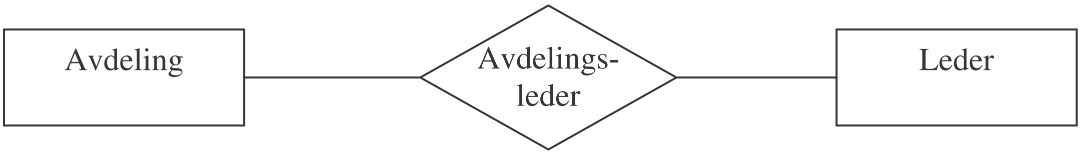
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En avdeling er en organisasjonsenhet som har en leveringsavtale med minst en leverandør.

En leder er en ansatt som har avdelingslederansvaret i en avdeling.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Avdelingsleder.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdelingsleder i forhold til relasjonen Avdelingsleder.

E. 0,1

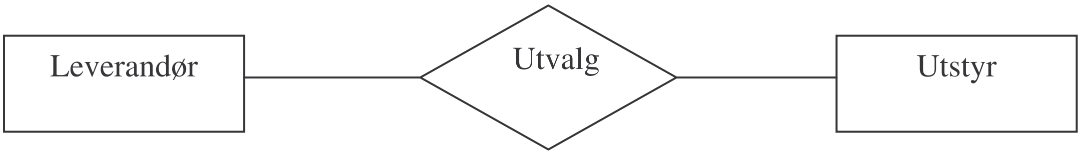
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En leverandør er et firma som har inngått en leveringsavtale med minst en avdeling.

Utstyr er en utstyrstype som inngår i sortimentet til en bestemt leverandør.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Utvalg.

I. 0,1

J. 0,N

K. 1,1

L. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Utstyr i forhold til relasjonen Utvalg.

I. 0,1

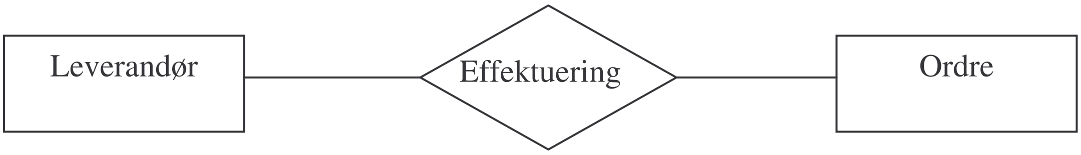
J. 0,N

K. 1,1

L. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En leverandør er et firma som har inngått en leveringsavtale med minst en avdeling.

En ordre er et formelt dokument for bestilling av utstyr som er datert og signert av en avdeling.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Effektuering.

M. 0,1

N. 0,N

O. 1,1

P. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ordre i forhold til relasjonen Effektuering.

M. 0,1

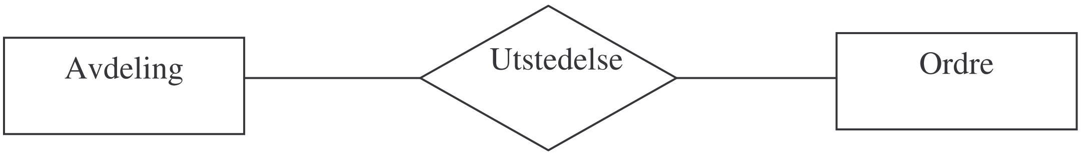
N. 0,N

O. 1,1

P. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En avdeling er en organisasjonsenhet som har en leveringsavtale med minst en leverandør.

En ordre er et formelt dokument for bestilling av utstyr som er datert og signert av en avdeling.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Utstedelse.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker	1	2	3	4	5	6	7	Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ordre i forhold til relasjonen Utstedelse.

E. 0,1

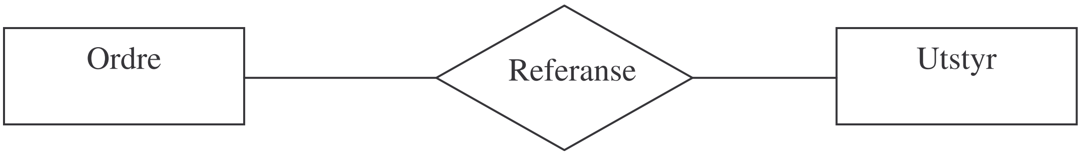
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker	1	2	3	4	5	6	7	Helt sikker



En ordre er et formelt dokument for bestilling av utstyr som er datert og signert av en avdeling.

Utstyr er en utstyrstype som inngår i sortimentet til en fast leverandør.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Ordre i forhold til relasjonen Referanse.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Utstyr i forhold til relasjonen Referanse.

E. 0,1

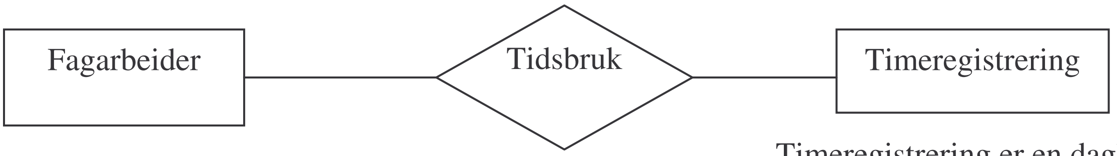
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En fagarbeider er en ansatt som har fagbrev i ett eller flere fag og som jobber fast på prosjekter.

Timeregistrering er en daglig oversikt over antall timer som en fagarbeider har jobbet med en arbeidsoppgave på et prosjekt.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Fagarbeider i forhold til relasjonen Tidsbruk.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Timeregistrering i forhold til relasjonen Tidsbruk.

E. 0,1

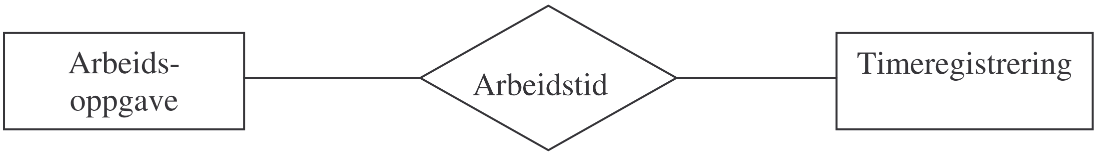
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



En arbeidsoppgave er en oppgave som det kan føres timer for.

Timeregistrering er en daglig oversikt over antall timer som en fagarbeider har jobbet med en arbeidsoppgave på et prosjekt.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Arbeidsoppgave i forhold til relasjonen Arbeidstid.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Timeregistrering i forhold til relasjonen Arbeidstid.

E. 0,1

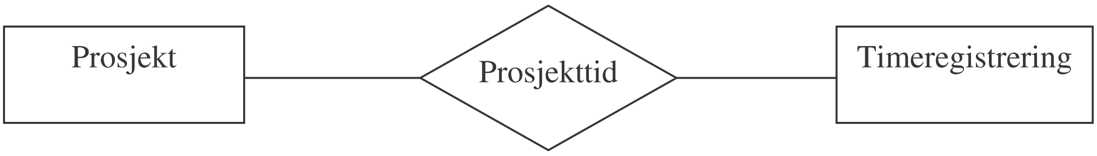
F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker



Et prosjekt er et oppdrag av en gitt varighet og med et gitt budsjett og en fast bemanning.

Timeregistrering er en daglig oversikt over antall timer som en fagarbeider har jobbet med en arbeidsoppgave på et prosjekt.

1. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Prosjekt i forhold til relasjonen Prosjekttid.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Timeregistrering i forhold til relasjonen Prosjekttid.

E. 0,1

F. 0,N

G. 1,1

H. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker 1 2 3 4 5 6 7 Helt sikker

Spørreskjema for demografiske data

1. Hvor erfaren er du med datamodellering?

Ikke erfaren 1 2 3 4 5 Erfaren

2. I hvilken grad er du fortrolig med ER-modellen?

Ikke fortrolig 1 2 3 4 5 Fortrolig

3. I hvilken grad har du erfaring med ER-modellering?

Ingen erfaring 1 2 3 4 5 Mye erfaring

4. I hvilken grad er du fortrolig med den syntaksen som er benyttet i testen?

Ikke fortrolig 1 2 3 4 5 Fortrolig

5. Hvor mange semester har du studert ved UiB? ____ semester inkludert inneværende semester.

6. Dersom du har relevant praksis i forhold til studiet, oppgi antall praksisår: ____ år.

7. Hvor gammel er du? ____ år.

8. Kjønn

Mann □

Kvinne □

Appendix B

Lecture held between experiment 1 and experiment 2

Arne Sølvberg

An information system is a *body of signs*, and the associated processes for storing and transforming the signs, and for exchanging signs with the exterior of the information system.

John Mylopoulos

Information modelling is concerned with the construction of computer-based *symbol structures*, which model some part of the real world.

Eksempler:

01102002	Europeisk format =	ddmmyyyy = 1 okt. 2002.
	US format =	mmddyyyy = 10 januar 2002.

Student En student er en person som har betalt semesteravgiften inneværende semester.

En student er en person som studerer eller har studert ved UiB.

Student er en rabatt som gis til personer som kan framvise studentbevis.

James Martin & James J. Odell

Object-oriented analysis is not an approach that models reality. Instead, it models the way people *understand and process* reality – through the concepts they acquire.

David Kroenke

A database does not model reality or a portion thereof. Instead, a database is a model of the *users' model*.

Grunnlagsdata - hva er det vi observerer?

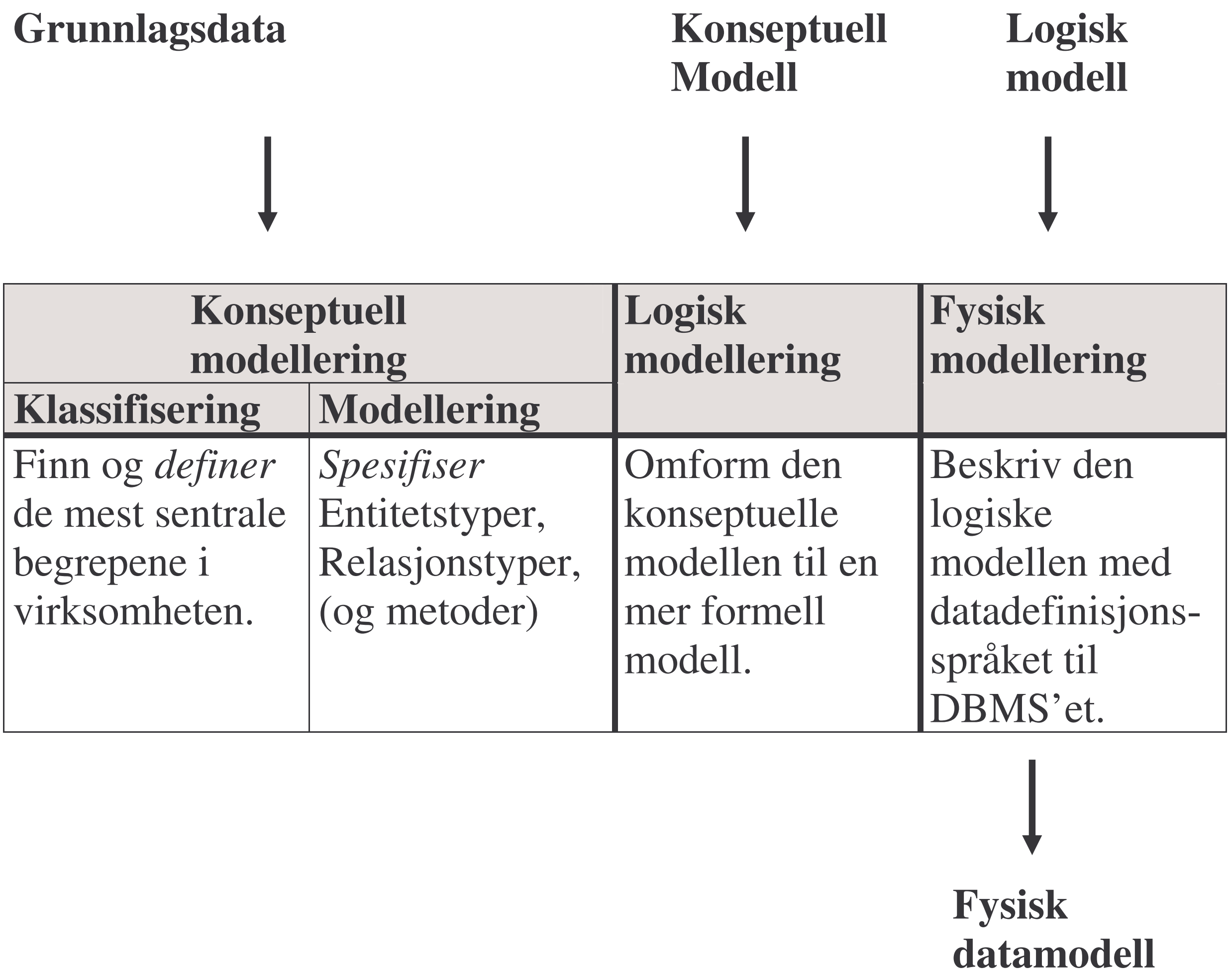
Intervju med brukere

- Problemer/mål/visjoner
- Informasjonsbehov
- Transaksjonsbehov

Dokumentstudier

- Skjema
- Journaler
- Rapporter
- Datalister
- Brosjyrer
- Regneark
- Skisser, tegninger
- Reglementer
- Systemdokumentasjon
- Faglitteratur

Datamodellering



Grunnlagsdata

Konseptuell Modell

Logisk modell

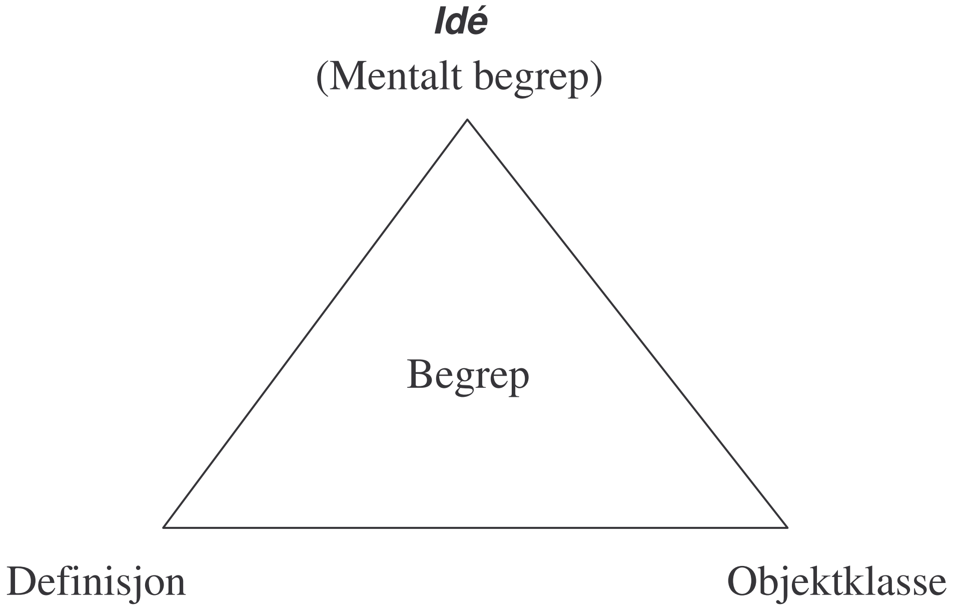
Konseptuell modellering		**Logisk modellering**	**Fysisk modellering**
Klassifisering	**Modellering**		
Finn og *definer* de mest sentrale begrepene i virksomheten.	*Spesifiser* Entitetstyper, Relasjonstyper, (og metoder)	Omform den konseptuelle modellen til en mer formell modell.	Beskriv den logiske modellen med datadefinisjonsspråket til DBMS'et.

Fysisk datamodell

Begrep

Et begrep er en ide eller forestilling som vi deler med andre og som refererer til visse objekter i vår bevissthet.

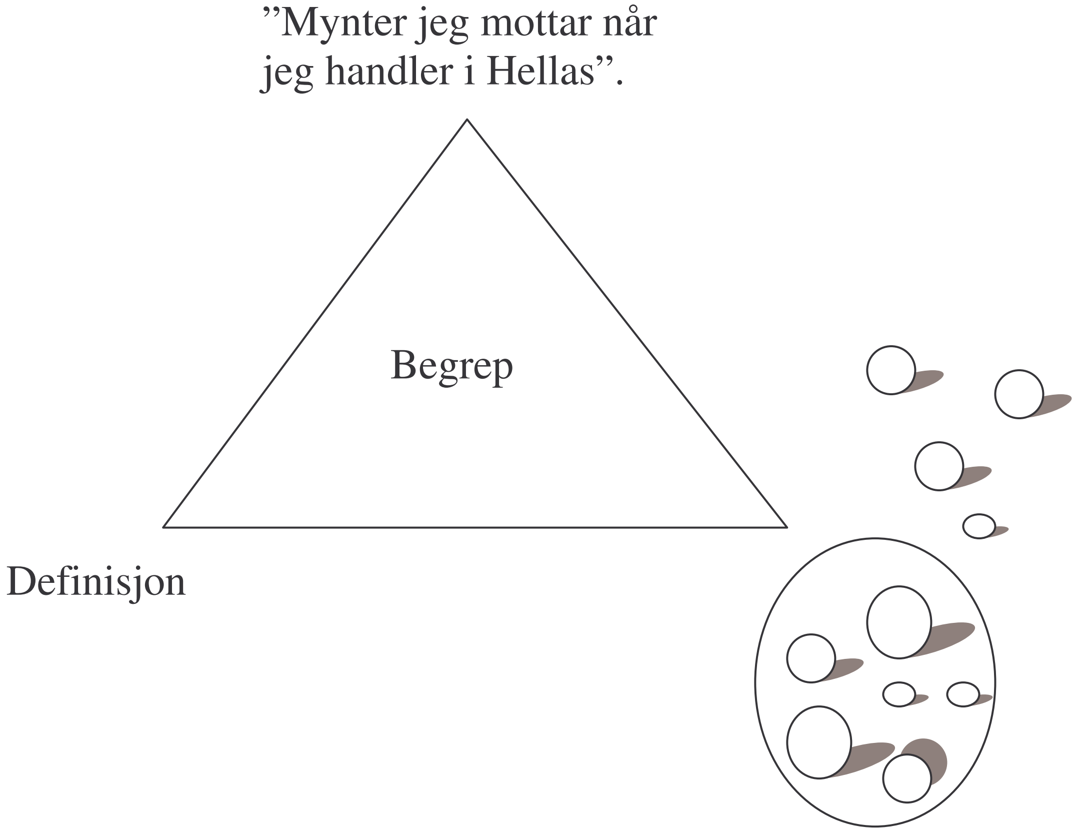
Et begrep er assosiert med en ide, en definisjon og et sett med objekter som faller inn under definisjonen.

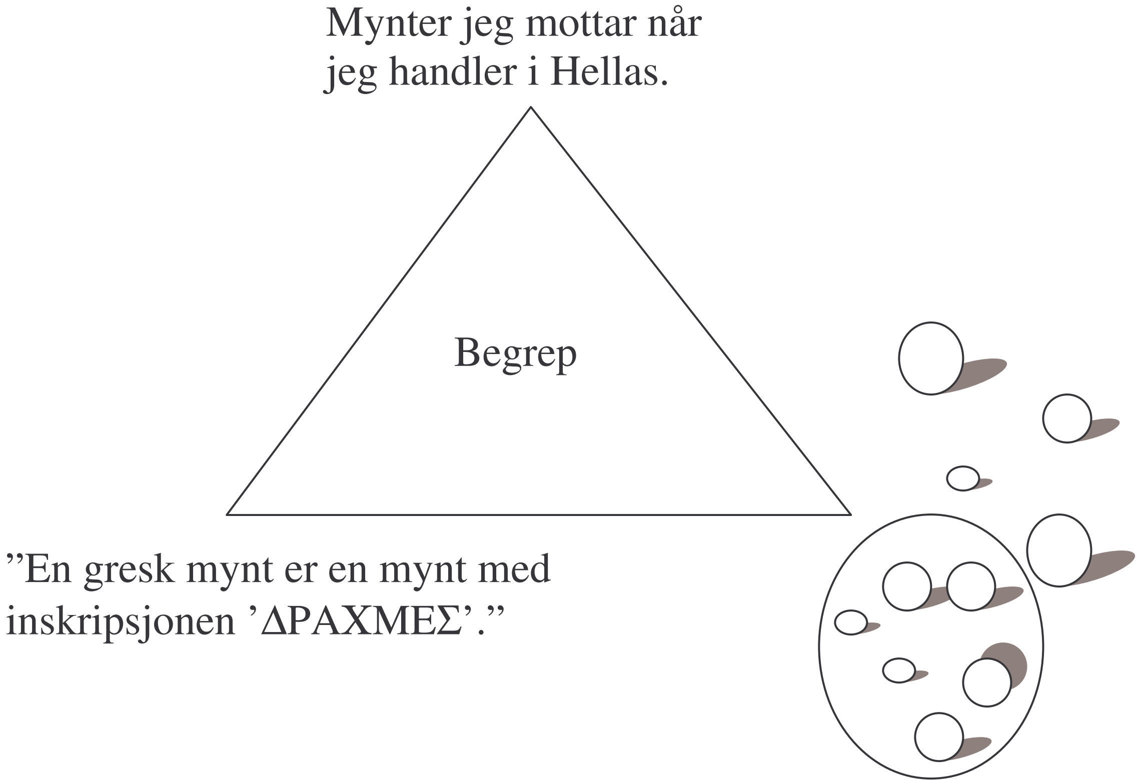


Idé: En subjektiv og privat test som brukes til å bestemme om et objekt omfattes av begrepet eller ikke.

Definisjon: En formell konkretisering av et mentalt begrep, som gjør begrepet tilgjengelig for andre.

Objektklasse: En logisk samling med objekter som tilfredsstiller betingelsen for medlemskap i klassen.

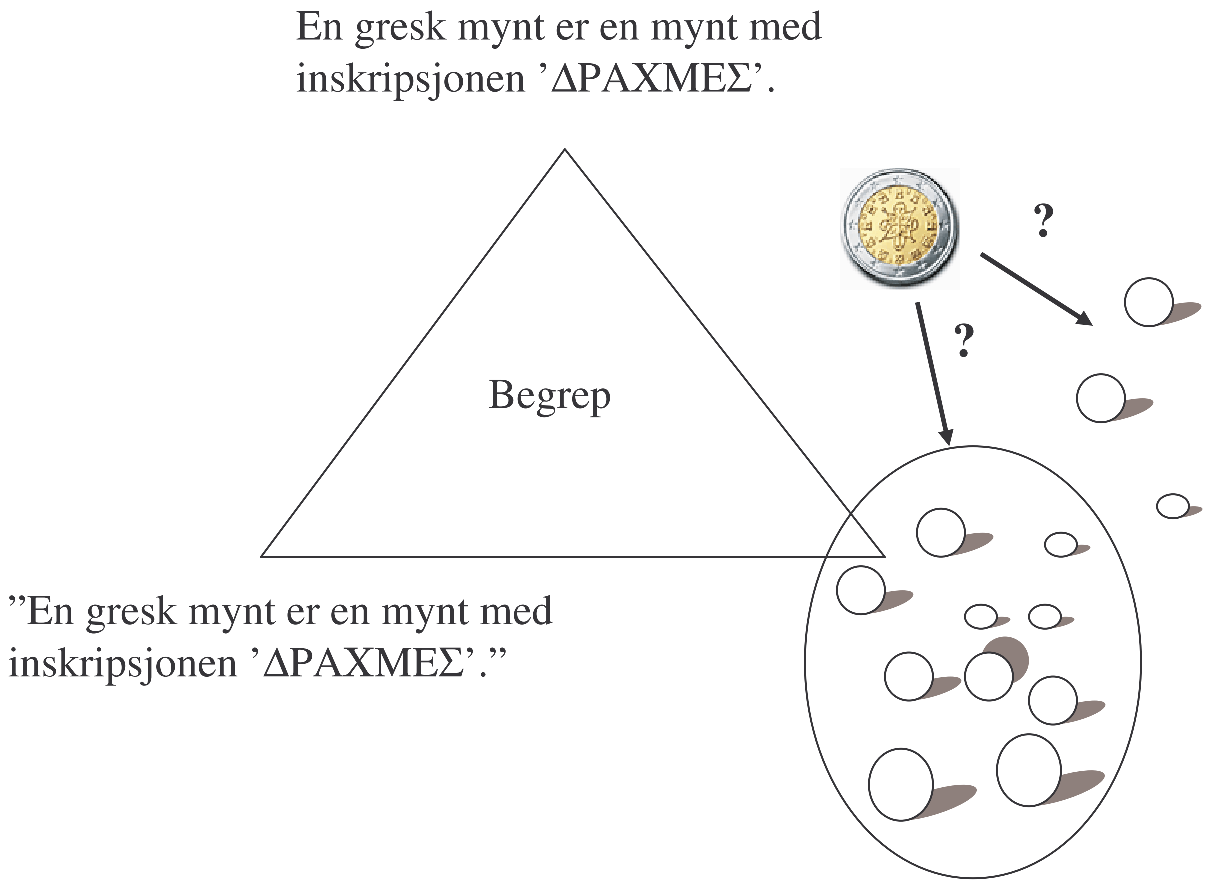

”Mynter jeg mottar når
jeg handler i Hellas”.
Begrep
Definisjon



Mynter jeg mottar når jeg handler i Hellas.

Begrep

”En gresk mynt er en mynt med inskripsjonen ’ΔΡΑΧΜΕΣ’.”


En gresk mynt er en mynt med inskripsjonen 'ΔΡΑΧΜΕΣ'.
Begrep
?
?
"En gresk mynt er en mynt med inskripsjonen 'ΔΡΑΧΜΕΣ'."

Klassifisering

1. Klassifisering (kategorisering) betyr å konstruere mentale begrep.

Mentalt begrep: En privat og subjektiv test som kan benyttes til å avgjøre om noe omfattes av begrepet eller ikke.

Eksempel:

- Mynter jeg mottar når jeg handler i Hellas.
- Mynter med inskripsjonen "ΔΡΑΧΜΕΣ".

2. Klassifisering betyr å konstruere begrepsdefinisjoner.

En begrepsdefinisjon består av en term som navngir begrepet, sammen med en definisjon som refererer til overordnet begrep og til karakteristika som skiller begrepet fra andre sideordnete begreper.

Eksempel:

Term ↓ Overordnet begrep ↓ Karakteristiske kjennetegn ↓

En gresk mynt er **en mynt** som **har inskripsjonen ‘ΔΡΑΧΜΕΣ’**

Venus er **en planet** som **….**

Venus er **en italiensk gud** som **….**

Venus er **en eksepsjonelt vakker kvinne** som **….**

3. Klassifisering brukes til å betegne det begrepssystemet som resulterer fra klassifikasjonsprosessen.

Et begrepssystem er en logisk konsistent samling med begrepsdefinisjoner.

Eksempel:

Et betalingsmiddel er en …

En mynt er et betalingsmiddel som …

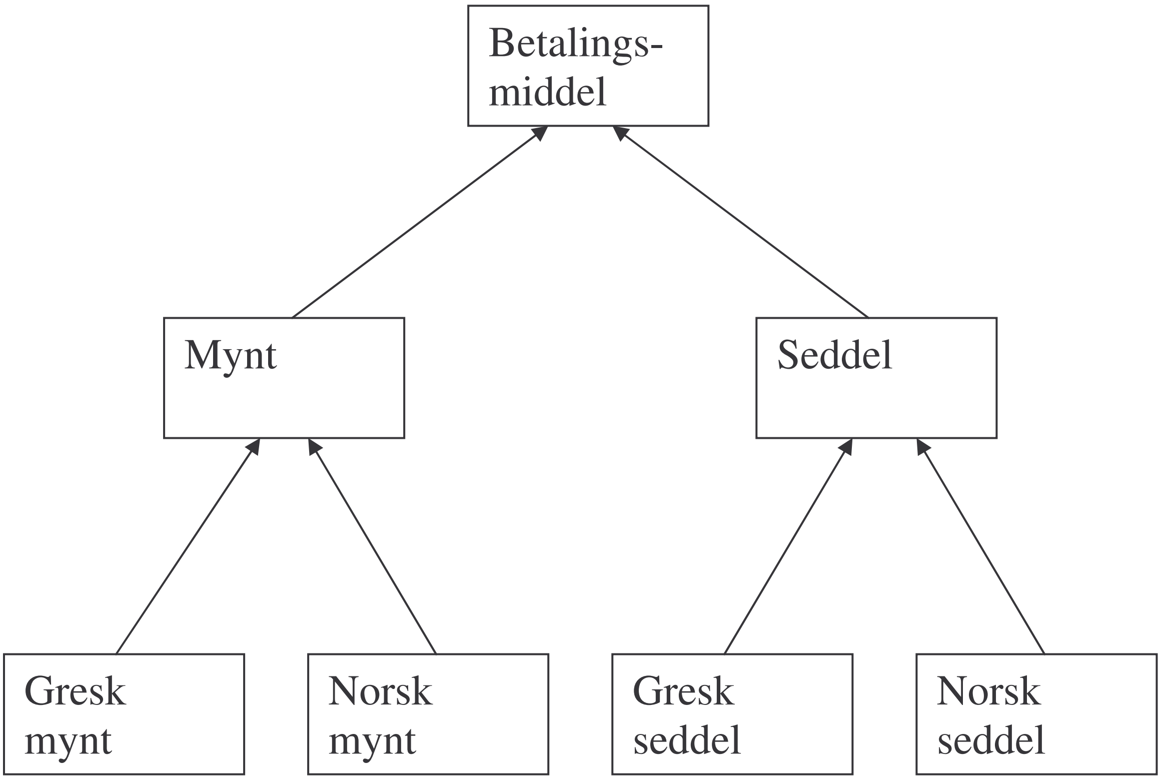
En gresk mynt er en mynt som har inskripsjonen **‘**ΔΡΑΧΜΕΣ**’**

En norsk mynt er en mynt som har inskripsjonen ’KRONER’ eller ’ØRE’

En seddel er et betalingsmiddel som er …

En gresk seddel er en seddel som har inskripsjonen **‘**ΔΡΑΧΜΕΣ**’**

En norsk seddel er en seddel som har inskripsjonen ’KRONER’


Betalings-middel
Mynt
Seddel
Gresk mynt
Norsk mynt
Gresk seddel
Norsk seddel

4. Klassifisering brukes til å betegne den vurderingen som utøves for å bestemme om et objekt faller inn under begrepet eller ikke.

Eksempel:

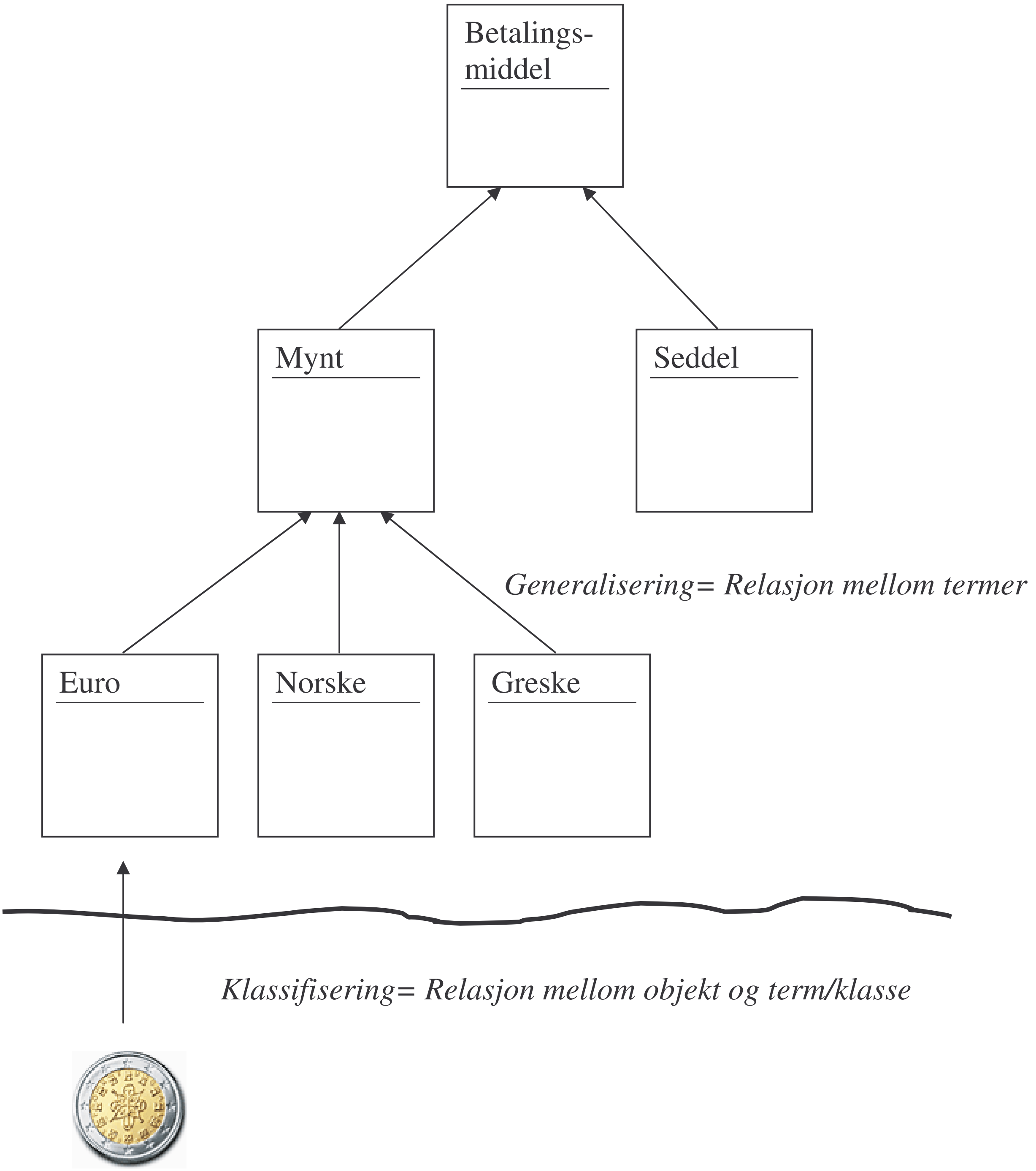
Er dette en gresk mynt?

Ja dersom den tilfredsstiller medlemskapskriteriet for greske mynter:

En gresk mynt er en mynt som har inskripsjonen **'ΔΡΑΧΜΕΣ'**

Et annet navn som brukes på denne vurderingsprosessen er **identifisering**.

5. Klassifisering brukes som en betegnelse på relasjonen mellom objekter og deres respektive klasser.



Generalisering= Relasjon mellom termer

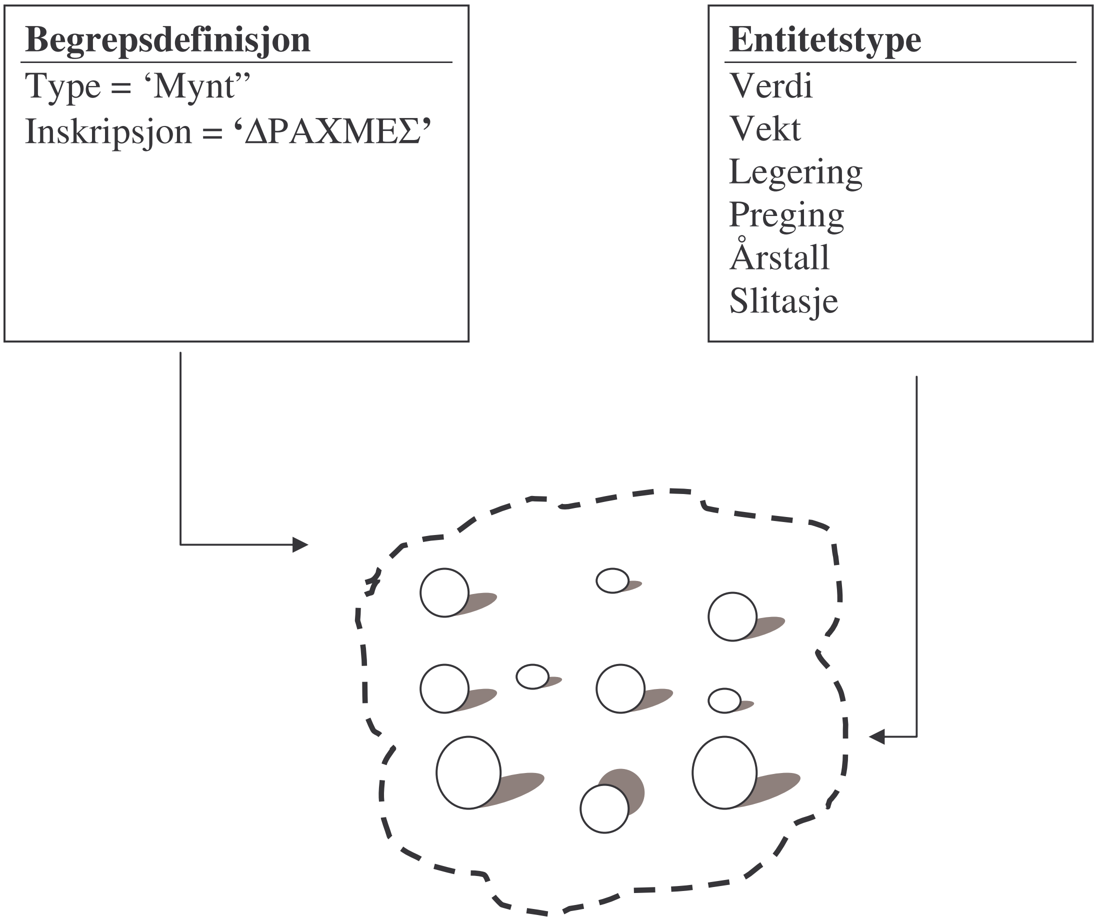
Klassifisering= Relasjon mellom objekt og term/klasse

Klassifisering og konseptuell modellering

Konseptuell modellering består av klassifisering og modellering.

- **Klassifisering** går ut på å *definere* begreper.

- **Modellering** går ut på å *spesifisere* entitetstyper og relasjonstyper.

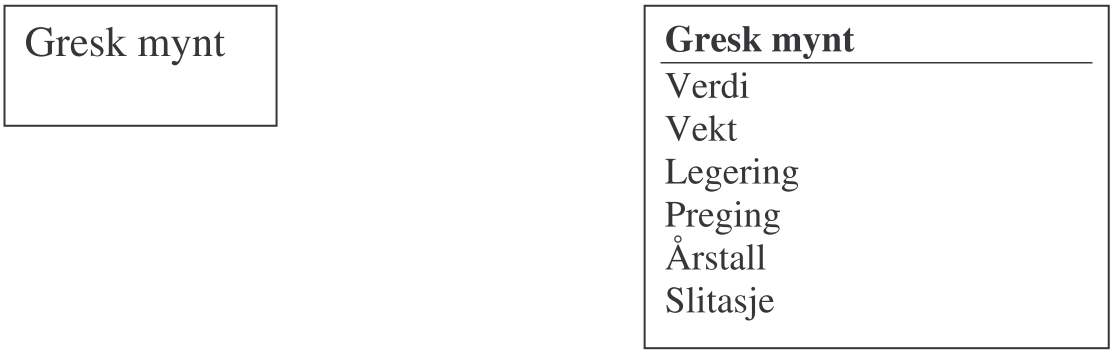


En **begrepsdefinisjon** refererer til en klasse og representerer betingelsene for medlemskap i klassen.

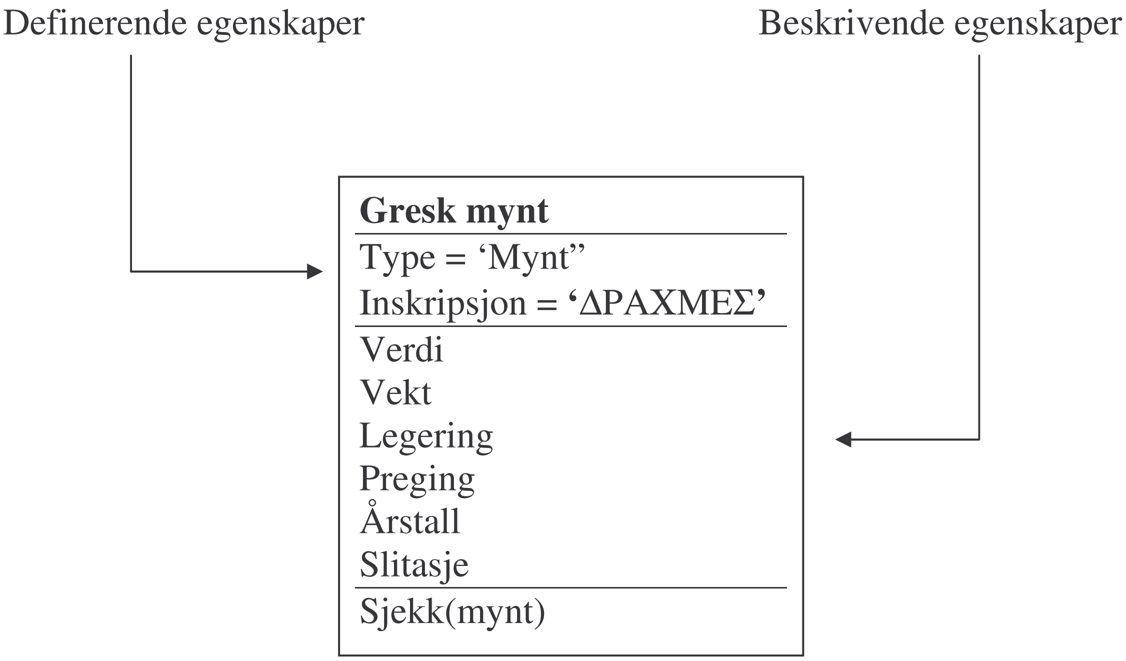
En **klasse** er en logisk samling med objekter som tilfredsstiller medlemskapsbetingelsene til klassen.

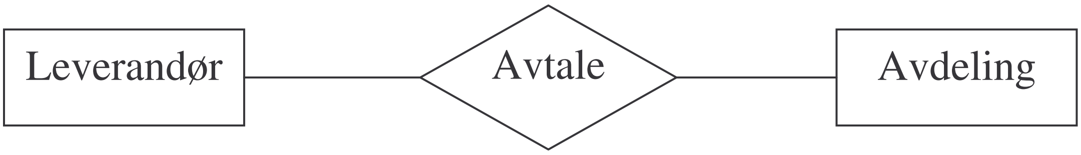
En **type** refererer til en klasse og representerer de attributtene som brukes til å beskrive medlemmene i klassen.

Entitetstyper i tradisjonell konseptuell modellering.



En mer komplett entitetstype:





1. *Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Avtale.*

   A. 0,1

   B. 0,N

   C. 1,1

   D. 1,N

2. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker	1	2	3	4	5	6	7	Helt sikker

3. *Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Avtale.*

   A. 0,1

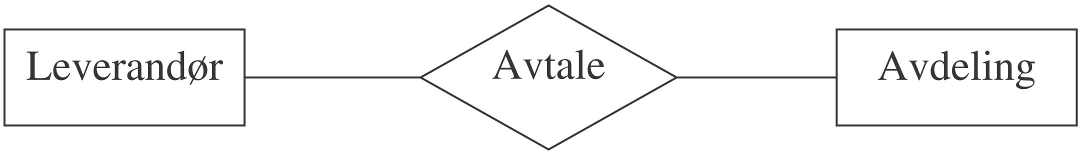
   B. 0,N

   C. 1,1

   D. 1,N

4. Hvor sikker er du med hensyn til det valget du har gjort.

Helt usikker	1	2	3	4	5	6	7	Helt sikker



En leverandør er et firma som har inngått en leveringsavtale med minst en avdeling.

En avdeling er en organisasjonsenhet som har en leveringsavtale med minst en leverandør.

1. *Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Leverandør i forhold til relasjonen Avtale.*

   A. 0,1

   B. 0,N

   C. 1,1

   D. 1,N

2. *Hvor sikker er du med hensyn til det valget du har gjort.*

Helt usikker 1 2 3 4 5 6 7 Helt sikker

3. *Sett en sirkel rundt det alternativet som du mener representerer de mest korrekte relasjonsbetingelsene for Avdeling i forhold til relasjonen Avtale.*

   A. 0,1

   B. 0,N

   C. 1,1

   D. 1,N

4. *Hvor sikker er du med hensyn til det valget du har gjort.*

Helt usikker 1 2 3 4 5 6 7 Helt sikker

Påstand

Konseptuell modellering består av to atskilte prosesser:

- Klassifikasjonsprosessen som angår *definering* av begreper.
- Modelleringsprosessen som angår *spesifisering* av entitetstyper, relasjonstyper og attributt-typer.

Definering av begrepene er en forutsetning for å identifisere det som skal beskrives. Hvis vi ikke kan identifisere det som skal beskrives, har vi ingenting å beskrive.

Gevinster ved å introdusere klassifisering

1. Begrepsbruk

Klassifisering vil kunne resultere i et **komplett** og **logisk konsistent** begrepsapparat.

- Begrepsapparatet vil være **komplett** når det omfatter de begrepene som er nødvendige for å imøtekomme informasjonsbehovene i applikasjonen.

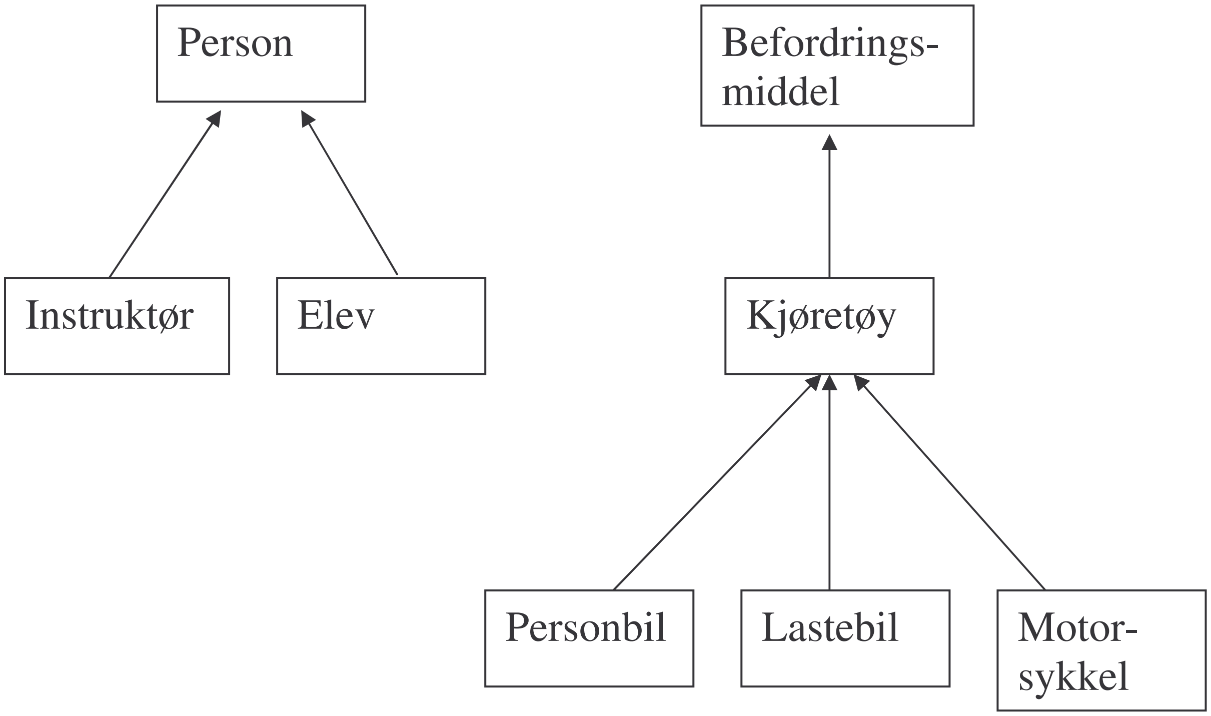
- Begrepsapparatet vil være **logisk konsistent** når alle begreper er definert med intensjonelle definisjoner.

Dermed kan begrepsapparatet gjøres til gjenstand for en formell evaluering av kompletthet og logisk konsistens.

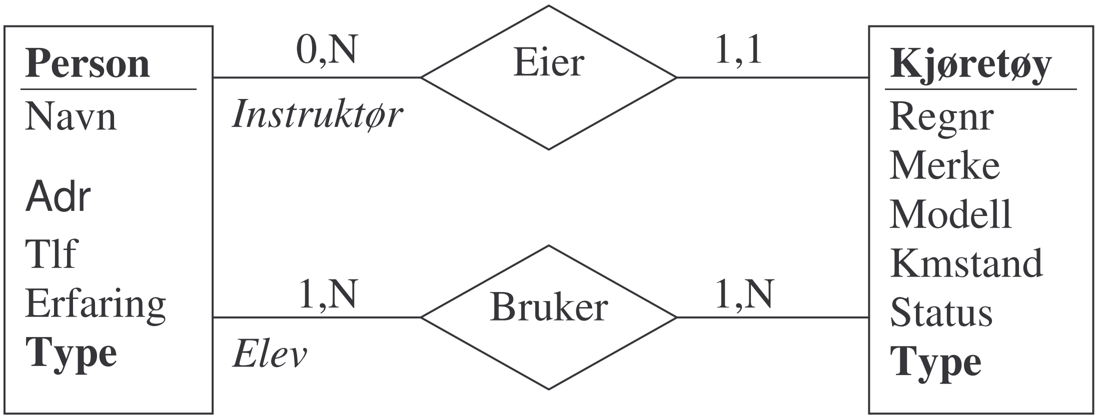
Begrepsapparatet vil videre bidra til identifisering, navngiving og definering av entitetstyper, relasjoner, roller og generaliseringshierarkier i den konseptuelle datamodellen.

Avvik i den konseptuelle modellen fra begreper og strukturer i begrepsapparatet bør gjøres rede for og begrunnes.

Begrepsstruktur



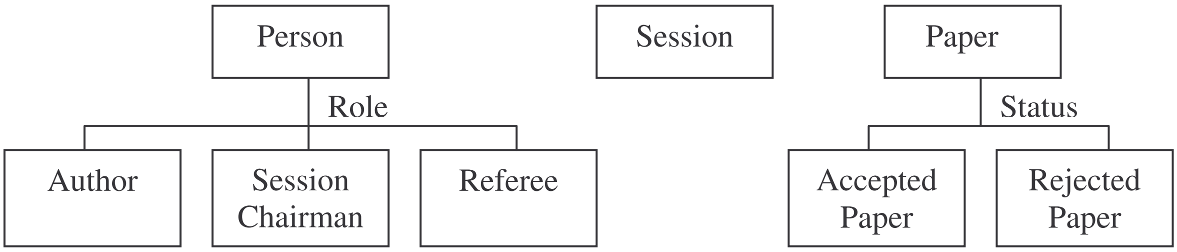
Konseptuell modell



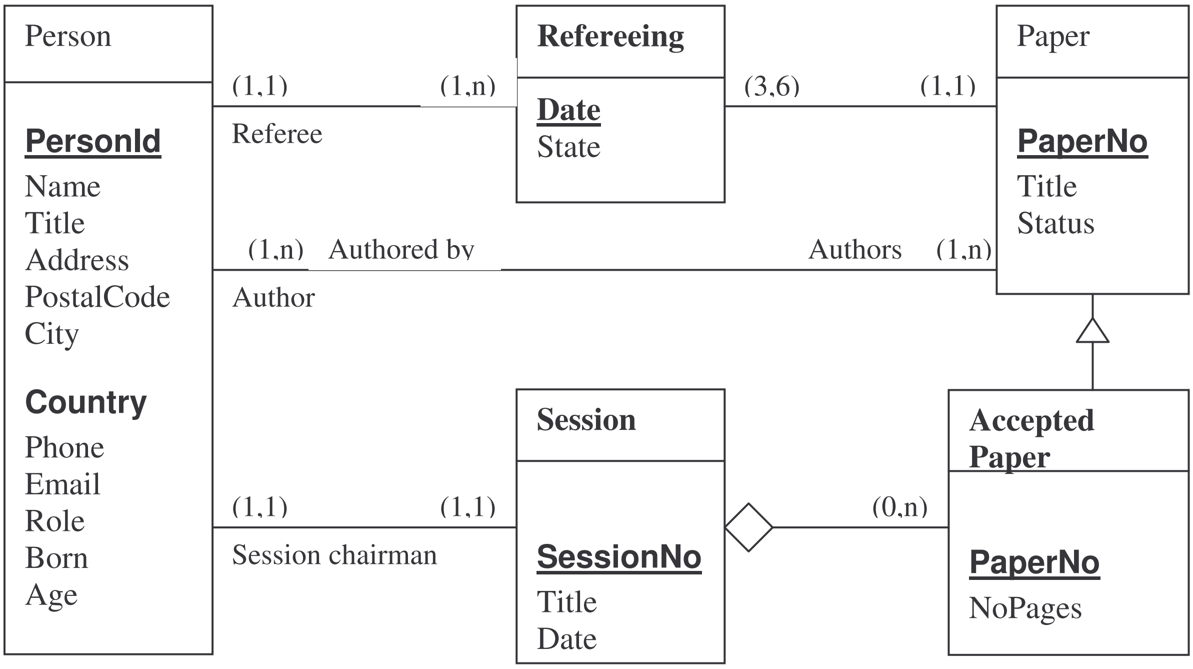
Person.type = {Instruktør | Elev}

Kjøretøy.type = {Personbil | Lastebil | Motorsykkel}

Example classification systems for a conference application



Example conceptual data model for a conference application

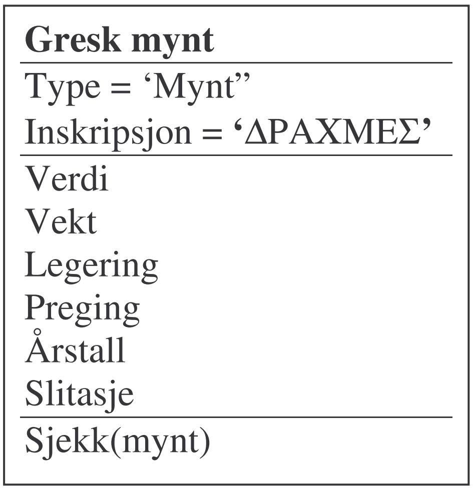


2. Dataintegritet.

Medlemskapsbetingelser gjør det mulig å kontrollere at objekter som registreres i databasen virkelig hører hjemme der.

- Å kontrollere at medlemskapskriteriene blir overholdt er et grunnleggende krav som bør formaliseres og kontrolleres av applikasjonen og ikke bare av brukeren.

- Medlemskapsbetingelser bør kunne testes ved hjelp av egne metoder.



Gresk mynt

Type = ‘Mynt”
Inskripsjon = ‘ΔΡΑΧΜΕΣ’

Verdi
Vekt
Legering
Preging
Årstall
Slitasje

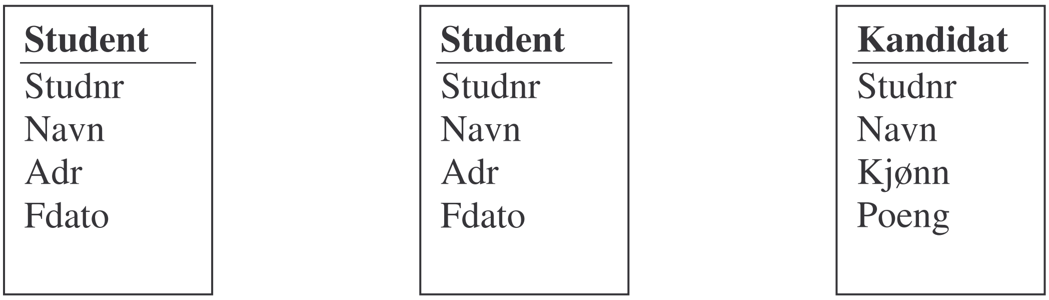
Sjekk(mynt)

3. Dataintegrering.

Ved forsøk på å integrere separate applikasjoner, er medlemskapsbetingelser den eneste muligheten til å fastslå om objektene i en klasse er av samme type som objektene i en annen klasse.

- Objekter kan beskrives på forskjellig vis i ulike applikasjoner. Selv om beskrivelsene er forskjellige kan objektene være de samme.

- Objekter kan beskrives på samme måte, men likevel være av forskjellige typer. Selv om beskrivelsene er like, kan objektene være forskjellige.



Forskningsspørsmål

1. Litteraturstudie

Studere hvordan klassifisering og relaterte begreper som *term, begrep, objekt, klasse, og medlemskapsbetingelser* blir beskrevet og brukt i lærebøker og artikler som omhandler konseptuell modellering.

2. Effekten av klassifisering

Studere effekten av medlemskapsbetingelser i forbindelse med:

- Tolking av konseptuelle modeller,
- Konstruksjon av konseptuelle modeller.

Appendix C

SPSS Output files from experiment 1 and 2

T-Test - Experiment 1

Group Statistics

	Treatment	N	Mean	Std. Deviation	Std. Error Mean
common sense	without membership conditions	16	10,69	2,152	,538
	with membership conditions	17	7,76	3,173	,769
mean confidence of all cases	without membership conditions	16	4,491	1,1895	,2974
	with membership conditions	17	4,395	1,4689	,3563
number of correct common sense answers with informative membership conditions	without membership conditions	16	5,69	1,493	,373
	with membership conditions	17	3,65	1,902	,461
mean confidence with informative mc	without membership conditions	16	4,3631	1,08963	,27241
	with membership conditions	17	4,3924	1,42907	,34660
number of correct answers with none-informative mc	without membership conditions	16	4,94	1,063	,266
	with membership conditions	17	3,88	1,933	,469
mean confidence with non-informative mc	without membership conditions	16	4,6338	1,36660	,34165
	with membership conditions	17	4,3776	1,56238	,37893

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
common sense	Equal variances assumed	1,215	,279	3,077	31	,004	2,92	,950	,986	4,860
	Equal variances not assumed			3,113	28,259	,004	2,92	,939	1,000	4,845
mean confidence of all cases	Equal variances assumed	1,373	,250	,205	31	,839	,096	,4671	-,8567	1,0485
	Equal variances not assumed			,207	30,348	,838	,096	,4641	-,8514	1,0432
number of correct common sense answers with informative membership conditions	Equal variances assumed	1,832	,186	3,413	31	,002	2,04	,598	,821	3,260
	Equal variances not assumed			3,439	30,063	,002	2,04	,593	,829	3,252
mean confidence with informative mc	Equal variances assumed	2,090	,158	-,066	31	,948	-,0292	,44450	-,93580	,87734
	Equal variances not assumed			-,066	29,760	,948	-,0292	,44084	-,92984	,87139
number of correct answers with none-informative mc	Equal variances assumed	4,115	,051	1,926	31	,063	1,06	,548	-,062	2,173
	Equal variances not assumed			1,958	25,160	,061	1,06	,539	-,054	2,164
mean confidence with non-informative mc	Equal variances assumed	,565	,458	,500	31	,621	,2561	,51234	-,78882	1,30103
	Equal variances not assumed			,502	30,845	,619	,2561	,51021	-,78469	1,29690

T-Test - Experiment 2

Group Statistics

	treatment	N	Mean	Std. Deviation	Std. Error Mean
common sense	Without membership conditions	14	12,86	2,685	,718
	with membership conditions	14	9,43	2,954	,789
mean confidence level for all cases	Without membership conditions	14	3,0914	1,69899	,45407
	with membership conditions	14	4,4907	1,54380	,41260
correct answers based on common sense and informative membership conditions	Without membership conditions	14	7,43	2,138	,571
	with membership conditions	14	4,79	1,805	,482
mean confidence for answers based on informative membership conditions	Without membership conditions	14	3,0800	1,68015	,44904
	with membership conditions	14	4,5814	1,58106	,42255
correct answers based on common sense and non-informative membership conditions	Without membership conditions	14	5,36	1,692	,452
	with membership conditions	14	4,64	1,946	,520
mean confidence for answers based on non-informative membership conditions.	Without membership conditions	14	3,1500	1,77155	,47347
	with membership conditions	14	4,3643	1,54053	,41172

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
common sense	Equal variances assumed	,591	,449	3,214	26	,003	3,43	1,067	1,236	5,621
	Equal variances not assumed			3,214	25,767	,004	3,43	1,067	1,235	5,622
mean confidence level for all cases	Equal variances assumed	,131	,720	-2,281	26	,031	-1,3993	,61353	-2,66042	-,13815
	Equal variances not assumed			-2,281	25,765	,031	-1,3993	,61353	-2,66098	-,13760
correct answers based on common sense and informative membership conditions	Equal variances assumed	,357	,555	3,534	26	,002	2,64	,748	1,106	4,180
	Equal variances not assumed			3,534	25,289	,002	2,64	,748	1,104	4,182
mean confidence for answers based on informative membership conditions	Equal variances assumed	,016	,900	-2,435	26	,022	-1,5014	,61659	-2,76886	-,23400
	Equal variances not assumed			-2,435	25,905	,022	-1,5014	,61659	-2,76908	-,23377
correct answers based on common sense and non-informative membership conditions	Equal variances assumed	,480	,495	1,037	26	,310	,71	,689	-,702	2,131
	Equal variances not assumed			1,037	25,508	,310	,71	,689	-,704	2,132
mean confidence for answers based on non-informative membership conditions.	Equal variances assumed	,636	,432	-1,935	26	,064	-1,2143	,62745	-2,50402	,07545
	Equal variances not assumed			-1,935	25,508	,064	-1,2143	,62745	-2,50523	,07666

Appendix D

Instruments for the measurement of design tasks

Oppgave:

1. Lag en konseptuell modell på grunnlag av den informasjonen du har tilgang til i dette oppgavesettet.
2. Ta utgangspunkt i den konseptuelle modellen og lag deretter en logisk datamodell.
3. Løs oppgavene alene, uten å samarbeide med andre. Lever besvarelsene sammen med spørreskjemaet som er vedlagt oppgavesettet. Husk at alle punktene i spørreskjemaet må fylles ut.

Systembeskrivelse.

Et reisebyrå som har spesialisert seg på salg av storbyferier i Europa ønsker en database for å holde orden på blant annet kunder, hoteller, bestillinger, ledere, selgere, sekretærer og guider. For hver ansatt skal det registreres ansattnr, navn og adresse. Dersom en ansatt for tiden er gift med en annen ansatt, skal det lagres data om vielsesdato og hvem som er gift med hvem. Det skal imidlertid ikke registreres data om ekteskap dersom en av partene ikke er ansatt i reisebyrået. For ledere skal det registreres bonus og frynsegoder (for eksempel avis, telefon, firmabil). For selgere skal det registreres et entydig selgernummer og hvor mye de skal selge for i året (salgsmål). For sekretærer skal det registreres hvilken oppgave de har (for eksempel regnskap, sentralbord, informasjon) og hvilke språk de behersker. For guider skal det registreres mobiltelefon, email og hvilke kurs de har tatt. En guide må kunne guide i minst en storby og i en storby kan det være behov for flere guider. For hver storby er det registrert navn, land, severdigheter og turistattraksjoner.

Reisebyrået har flere filialer og for hver filial er det registreres navn, tlf, fax, email og postadresse. Hver filial har flere ansatte og en ansatt er kun ansatt ved en filial. En filial identifiseres med navn og styres av en leder (som bare kan styre en filial).

Hver filial har avtale med flere hotellkjeder. En hotellkjede kan ha avtale med flere filialer. En filial har kun en avtale med en hotellkjede. For hver avtale skal det registreres dato for når avtalen ble inngått eller sist ble revidert. For de hotellkjedene som det er inngått avtale med skal det registreres navn, adresse, tlf, fax, email og web-adresse.

En hotellkjede driver flere hoteller. Hvert hotell er lokalisert i en storby og hvert hotell består av mange rom. For hvert hotell registreres navn (som er entydig), adresse, tlf, standard og antall rom. For hvert rom registreres romnr, type, TV, dusj, wc.

En selger kan formidle mange hoteller til mange kunder. En kunde kan få formidlet mange hoteller fra mange selgere. Et hotell kan formidles til mange kunder fra mange selgere. Ved første gangs formidling skal hotellformidlingen registreres med attributtet *antall_formidlinger* = 1. Samtidig skal kunden registreres med kundenr, navn, adresse og telefon. Ved eventuelle senere formidlinger av samme hotell til samme kunde av samme selger, skal dette registreres ved å oppdatere attributtet *antall_formidlinger* med 1.

I tillegg til å registrere hotellformidlinger, skal det være mulig å booke flyreiser. En flyreise angår en kunde, en flight og en selger. En kunde kan bestille flere flighter av forskjellige selgere. En selger kan booke flere flighter til forskjellige kunder. En flyreise kan bare bookes av en selger. For hver flyreise må det registreres bookingdato. I databasen skal det finnes en oversikt av flighter og flyplasser. Hver flight er registrert med flightnummer, avreiseflyplass, ankomstflyplass, avreisedato og avreisetidspunkt. Hver flyplass er registrert med plasskode, plassnavn, land, nærmeste by, avstand til nærmeste by og transportmuligheter.

Begrepsapparat

Objektklasser

By = ikke nærmere definert.

- **Storby** = en by som ligger i Europa og som har mer enn 1 million innbyggere.

Forretningsvirksomhet = ikke nærmere definert.

- **Hotellkjede** = en forretningsvirksomhet som driver flere hoteller.

Lufthavn = ikke nærmere definert.

- **Flyplass =** en lufthavn for sivil flytrafikk.

Overnattingssted = ikke nærmere definert.

- **Hotel** = et overnattingssted som drives av en hotellkjede og som disponerer minst 25 rom.

Person = ikke nærmere definert.

- **Ansatt** = en person som er ansatt ved en av reisebyråets filialer.
  - **Leder** = en ansatt med siviløkonomutdannelse og minst fem års erfaring fra reiselivsnæringen..
  - **Selger** = en ansatt med utdannelse innen reiseliv.
  - **Sekretær** = en ansatt med utdannelse som sekretær som behersker minst ett fremmedspråk.
  - **Guide** = en ansatt som har bestått minst ett guidekurs.
- **Kunde** = en person som har blitt formidlet et hotell.

Reiseetappe = ikke nærmere definert.

- **Flight** = en reiseetappe med fly fra en flyplass til en annen.

Salgsenhet = ikke nærmere definert.

- **Filial** = En salgsenhet med egen leder og eget budsjett.

Værelse = ikke nærmere definert.

- **Rom** = et værelse som kan leies ut til kunder for en bestemt døgnpris.

Relasjonsklasser

Forbindelse = ikke nærmere definert

Relasjon en forbindelse mellom to eller flere entiteter.

Ansatt ved = en relasjon mellom Ansatt og Filial. En ansatt må være ansatt ved en filial, og en filial må ha minst en ansatt.

Ankomststed = en relasjon mellom Flyplass og Flight. En flyplass kan være ankomststed for mange flighter. En Flight må ha en flyplass som ankomssted.

Avreisested = en relasjon mellom Flyplass og Flight. En flyplass kan være avreisested for mange flighter. En flight må ha en flyplass som avreisested.

Avtale = en relasjon mellom en hotellkjede og en filial. En hotellkjede må ha en avtale med minst en filial. En filial må ha en avtale med minst en hotellkjede. En filial har kun en avtale med en hotellkjede.

Består av = en relasjon mellom Hotell og Rom. Et hotell må bestå av minst 25 rom. Et rom må høre til ett hotell.

Byguide = en relasjon mellom By og Guide. En by kan ha flere guider. En guide må kunne guide i minst en by.

Driver = en relasjon mellom Hotellkjede og Hotell. En hotellkjede må drive minst ett hotell. Et hotell må drives av en hotellkjede.

Ekteskap = en relasjon mellom to ansatte som er gift med hverandre. En ansatt kan være gift med en annen ansatt.

Flyreise = en relasjon mellom Kunde, Flight og Selger. En kunde kan bestille flere flighter fra forskjellige selgere. En selger kan booke flere flighter til forskjellige kunder. En flight kan bookes til flere kunder av forskjellige selgere. En flyreise kan bare bookes av en selger.

Hotellformidling = en relasjon mellom Selger, Kunde og Hotell. En selger kan formidle mange hoteller til mange kunder. En kunde kan få formidlet mange hoteller fra mange selgere. Et hotell kan formidles til mange kunder fra mange selgere.

Leder for = en relasjon mellom Leder og Filial. En leder må være leder for en filial. En filial må ha en leder.

Lokalisert i = en relasjon mellom Hotell og Storby. Et hotell må være lokalisert i en storby. En storby må være lokasjonssted til minst ett hotell.

Demografiske data

1. Hvor erfaren er du med datamodellering?

Ikke erfaren 1 2 3 4 5 Erfaren

2. I hvilken grad har du erfaring med ER-modellering?

Ingen erfaring 1 2 3 4 5 Mye erfaring

3. Synes du oppgaven har vært lett eller vanskelig?

Lett 1 2 3 4 5 Vanskelig

4. I hvor stor grad mener du at den konseptuelle modellen du har laget gir et korrekt bilde av hospitalsystemet?

Liten grad 1 2 3 4 5 Stor grad

5. Hvor mange semester har du studert ved universitet/høgskole? ____ semester.

6. Dersom du har relevant praksis i forhold til kurset, oppgi antall praksisår: ____ år.

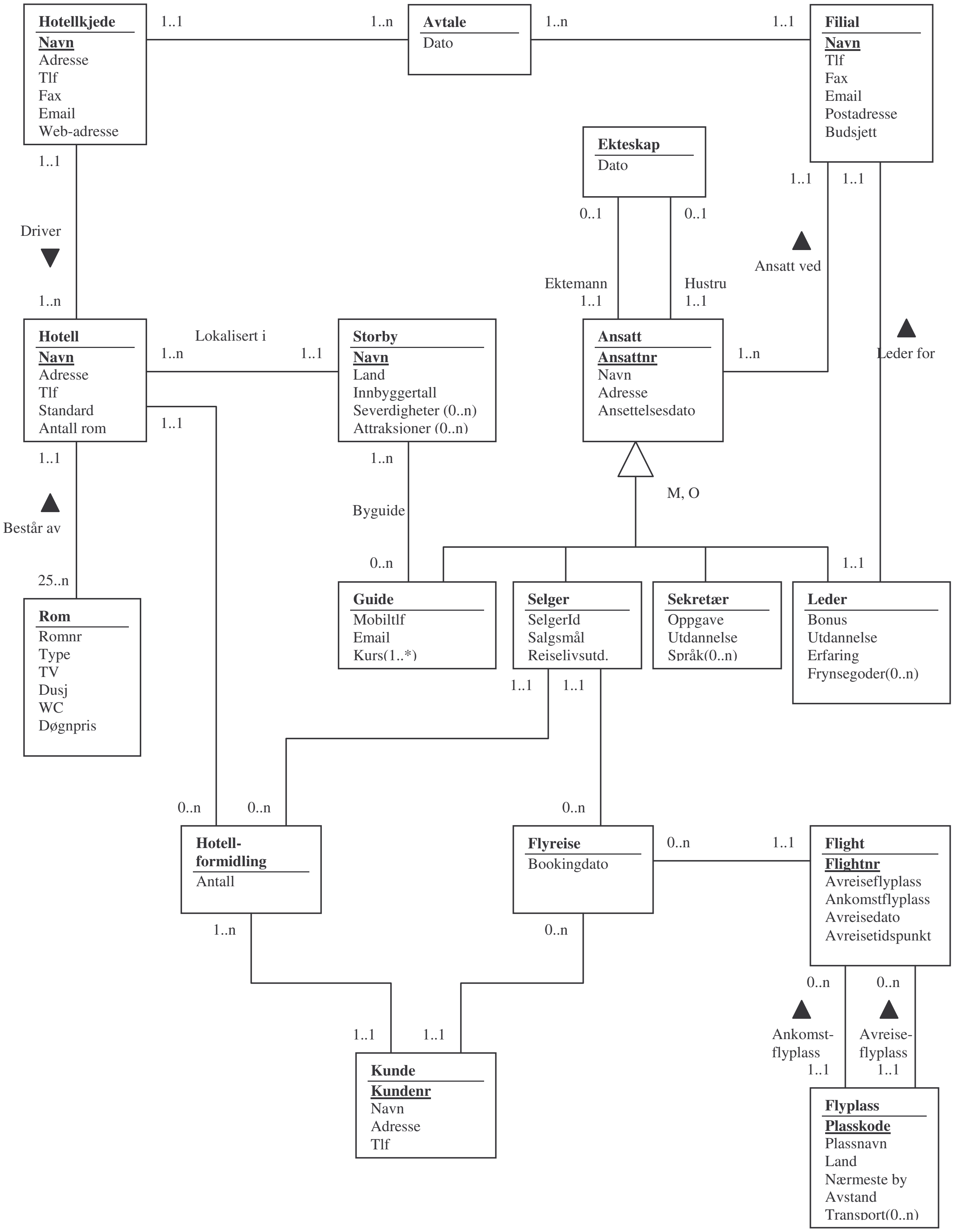
7. Hvor gammel er du? ____ år.

8. Kjønn

Mann □

Kvinne □

Konseptuell modell med entitetstyper og relasjonstyper.



Logisk modell:

Hotellkjede (**Navn**, Adresse, Tlf, Fax, Email, Web-adresse)

Storby (**Navn**, Land, Innbyggertall)

Severdigheter (***Severdighet,*** *Storby*)

Attraksjoner (***Attraksjon,*** Storby)

Hotell (**Navn**, Adresse, *Tlf*, Standard, AntallRom, *Hotellkjede*, *Storby*)

Rom (***Hotell,* Romnr,** Type, *TV*, Dusj, WC, Døgnpris)

Filial (**Navn**, Tlf, Fax, EMail, Postadresse, Budsjett, *Leder*)

Avtale (***Filial, Hotellkjede***, Dato)

Ansatt (**Ansattnr**, Navn, Adresse, Ansettelsesdato, *Filial*, Type)

Ekteskap (***Ektemann*,** *Hustru*, Dato)

Guide (***Ansattnr***, MobilTlf, Email)

Guidekurs (***Ansattnr*, Kurs**)

Selger (***Ansattnr***, SelgerId, Salgsmål, Reiselivsutdanning)

Sekretær (***Ansattnr***, Oppgave, Utdannelse)

Språk (***Ansattnr***, **Språk**)

Leder (**Ansattnr**, Bonus, Utdannelse, Erfaring)

Frynsegoder (***Ansattnr*, Gode**)

Byguide (***Ansattnr*, *Storby***)

Hotellformidling (***Hotell, Selger, Kunde,*** Antall_formidlinger)

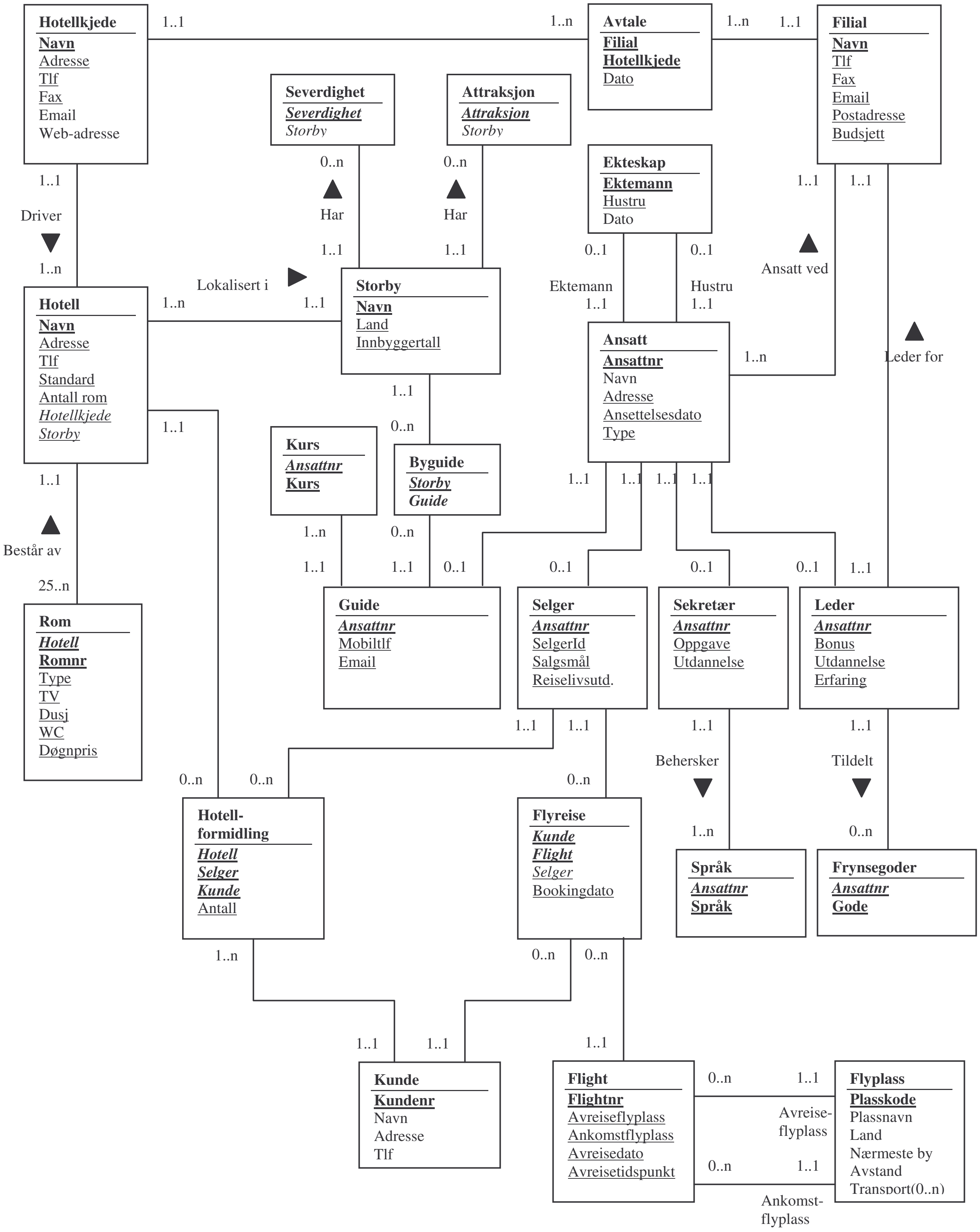
Flyplass (**Plasskode**, Plassnavn, Land, NærmesteBy, Avstand)

Transportmiddel (***Plasskode*, Transportmiddel**)

Flight (**Flightnr**, *Avreiseflyplass*, *Ankomstflyplass*, Avreisedato, Avreisetidspunkt)

Flyreise (***Kunde, Flight***, *Selger*, Bookingdato)

Skisse av logisk modell.



Oppgave:

4. Lag en konseptuell modell på grunnlag av den informasjonen du har tilgang til i dette oppgavesettet.

5. Ta utgangspunkt i den konseptuelle modellen og lag deretter en logisk datamodell.

6. Gjør egne forutsetninger i den grad du finner dette nødvendig og redegjør for disse både når det gjelder den konseptuelle og den logiske modellen.

Systembeskrivelse.

Et sykehus ønsker en database for å holde orden på data om leger, pleiere, sekretærer og vedlikeholdspersonell. For hver ansatt skal det registreres personnummer, navn og adresse. Dersom en ansatt for tiden er gift med en annen ansatt, skal det lagres data om vielsesdato og hvem som er gift med hvem. Det skal imidlertid ikke registreres data om ekteskap dersom en av partene ikke er ansatt på sykehuset. Hver lege på sykehuset har et entydig nummer og en eller flere spesialiteter. Det er nødvendig å lagre utdannelsen til hver pleier (for eksempel, hjelpepleier, sykepleier, etc.). Vedlikeholdspersonell kan ha flere arbeidsoppgaver (for eksempel renholdsarbeider, portør, etc.). En sekretær kan beherske flere tekstbehandlingsprogrammer.

Hver ansatt tilhører en avdeling. En avdeling identifiseres med navn. En avdeling ledes av en lege (som bare kan lede en avdeling). Det må registreres hvor mange sengeplasser som finnes i hver avdeling.

For hver pasient som blir innlagt på sykehuset skal det lagres data om personnummer, navn, kjønn, fødselsdato og innleggelsesdato. En pasient er innlagt på en avdeling. I innleggelsesperioden kan en pasient få stilt diagnoser for flere lidelser. Hver lidelse er karakterisert med en kode og et navn. For hver diagnose som blir stilt, skal det registreres dato og hvilken lege som har stilt diagnosen. En diagnose stilles av en lege.

I løpet av innleggelsesperioden, vil hver pasient gå gjennom flere undersøkelser. (Samme undersøkelse kan utføres på mange pasienter), og hun/han kan få ulike medikamenter. Resultatene fra hver undersøkelse som en pasient tar skal lagres. En undersøkelse blir identifisert med en testkode. En undersøkelse har også en beskrivelse og en pris. Hvert medikament har en medikamentkode, type (for eksempel tablett, væske, krem, etc.) og mengde i hver pakke. En pasient kan motta mer enn ett medikament. Det er viktig å lagre antall enheter en pasient bruker daglig av hvert medikament.

Medikamenter leveres til sykehuset som en leveranse. En leveranse mottas fra en leverandør (en leverandør kan ha et stort antall leveranser). Hver leveranse har et identitetsnummer og en leveransedato. En leveranse kan bestå av mange medikamenter som skal til ulike avdelinger på sykehuset. Det kan være at ett medikament skal til flere avdelinger. For hver leveranse må det registreres hvor stor mengde av hvert medikament en avdeling mottar. For hver leverandør skal det lagres navn og adresse. (Du kan anta at en leverandør er identifisert ved navnet).

Begrepsapparat

Objektklasser

Person = ikke nærmere definert.

- **Ansatt** = en person som er ansatt ved sykehuset.
    - **Lege** = en ansatt med utdannelse som lege og med spesialisering innen ett eller flere områder.
    - **Pleier** = en ansatt med utdannelse som sykepleier eller hjelpepleier.
    - **Sekretær** = en ansatt med utdannelse som sekretær og minst 3 års relevant praksis.
    - **Vedlikeholdspersonell** = en ansatt med teknisk eller praktisk utdannelse.
- **Pasient** = en person som er innlagt på en avdeling.

Organisasjonsenhet = ikke nærmere definert.

- **Avdeling** = En organisasjonsenhet med egen avdelingsleder, eget budsjett og et fast antall senger.

Medisinsk betegnelse = ikke nærmere definert.

- **Lidelse** = en medisinsk betegnelse på en skade eller sykdom som er klassifisert i henhold til icd10-standarden.

Medisinsk test = ikke nærmere definert.

- **Undersøkelse** = en medisinsk test som utføres på pasienter for å kartlegge pasientens kliniske tilstand og som er klassifisert i henhold til Norsk Klinisk Test Standard.

Preparat = ikke nærmere definert.

- **Medikament** = et preparat som brukes til medisinsk behandling av pasienter og som er klassifisert i henhold til Norsk Medikament Standard.

Firma = ikke nærmere definert.

- **Leverandør** = et firma som leverer medikamenter til sykehuset.

Postsending = ikke nærmere definert.

- **Leveranse** = en postsending med medikamenter som skal fordeles til en eller flere avdelinger.

Relasjonsklasser

Forbindelse = ikke nærmere definert

Relasjon en forbindelse mellom to eller flere entiteter.

Ekteskap = en relasjon mellom to ansatte som er gift med hverandre. En ansatt kan være gift med en annen ansatt.

Tilhører = en relasjon mellom Ansatt og Avdeling. En ansatt må tilhøre en avdeling, og en avdeling må ha minst en ansatt.

Avdelingsleder = en relasjon mellom Lege og Avdeling. En lege kan være avdelingsleder for en avdeling. En avdeling må ha en avdelingsleder.

Innlagt på = en relasjon mellom Pasient og Avdeling. En pasient må være innlagt på en avdeling. En avdeling kan ha flere pasienter innlagt.

Diagnose = en relasjon mellom Pasient, Lege og Lidelse. En pasient kan ha mange lidelser. En lidelse kan plage mange pasienter. En lege kan stille mange diagnoser. En diagnose blir stilt av bare en lege.

Resultater = en relasjon mellom Pasient og Undersøkelse. En pasient kan ta flere undersøkelser. En undersøkelse kan utføres på flere pasienter.

Medisinering = en relasjon mellom Medikament og Pasient. Et medikament kan brukes på flere pasienter. En pasient kan behandles med flere medikamenter.

Leverer = en relasjon mellom Leverandør og Leveranse. En leverandør kan levere flere leveranser. En leveranse må være levert av en leverandør.

Del-leveranse = en relasjon mellom Leveranse, Avdeling og Medikament. En leveranse kan bestå av flere medikamenter som skal til flere avdelinger. En avdeling kan motta flere medikamenter i samme leveranse. Et medikament kan leveres til flere avdelinger i samme leveranse.

Demografiske data

1. Hvor erfaren er du med datamodellering?

Ikke erfaren 1 2 3 4 5 Erfaren

2. I hvilken grad har du erfaring med ER-modellering?

Ingen erfaring 1 2 3 4 5 Mye erfaring

3. Synes du oppgaven har vært lett eller vanskelig?

Lett 1 2 3 4 5 Vanskelig

4. I hvor stor grad mener du at den konseptuelle modellen du har laget gir et korrekt bilde av hospitalsystemet?

Lite grad 1 2 3 4 5 Stor grad

5. Hvor mange semester har du studert ved universitet/høgskole? ____ semester.

6. Dersom du har relevant praksis i forhold til kurset, oppgi antall praksisår: ____ år.

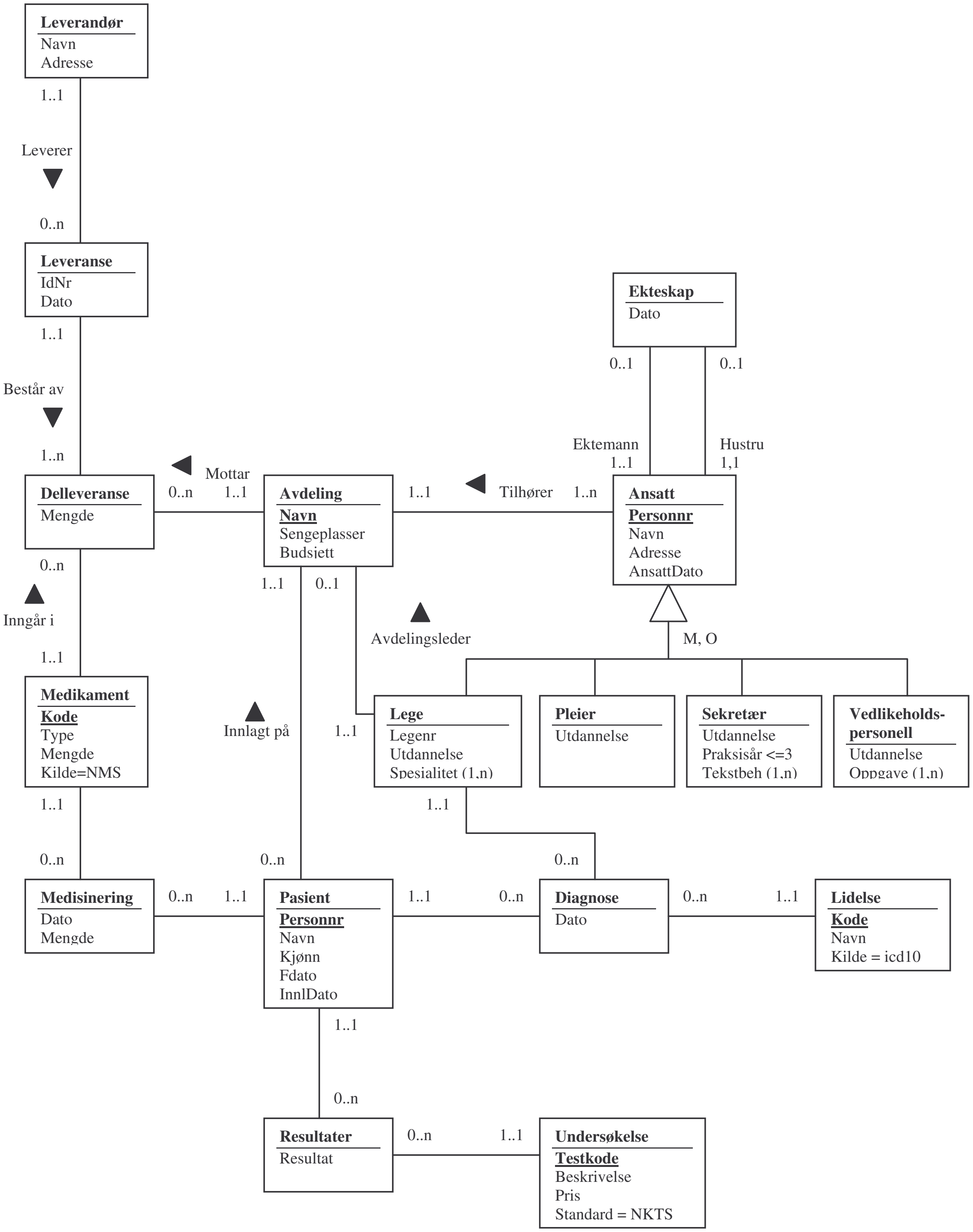
7. Hvor gammel er du? ____ år.

8. Kjønn

Mann □

Kvinne □

Konseptuell modell med entitetstyper og relasjonstyper.



Logisk modell:

Leverandør (**Navn**, Adresse)

Leveranse (**idNr**, Dato, *Leverandør*)

Avdeling (**Navn**, Sengeplasser, Budsjett, *Avdelingsleder*)

Medikament (**Kode**, Type, Mengde, Kilde)

Delleveranse (***Leveranse, Medikament, Avdeling***, mengde)

Ansatt (**Personnr**, Navn, Adresse, *Avdeling, Utdannelse, Ansattdato,* Type)

Ekteskap (***Ektemann*****,** *Hustru*, Dato)

Lege (***Personnr***, Legenr)

Spesialitet (***Personnr,*** **spesialitet**)

Sekretær (***Personnr***, Praksisår)

Tekstbehandling (***Personnr*****, tekstbehandling**)

Oppgaver (***Personnr*****, oppgave**)

Pasient (**Personnr**, Navn, Kjønn, Fdato, InnlDato, *Avdeling*)

Medisinering (***Personnr, MedikamentKode*****, Dato, Mengde**)

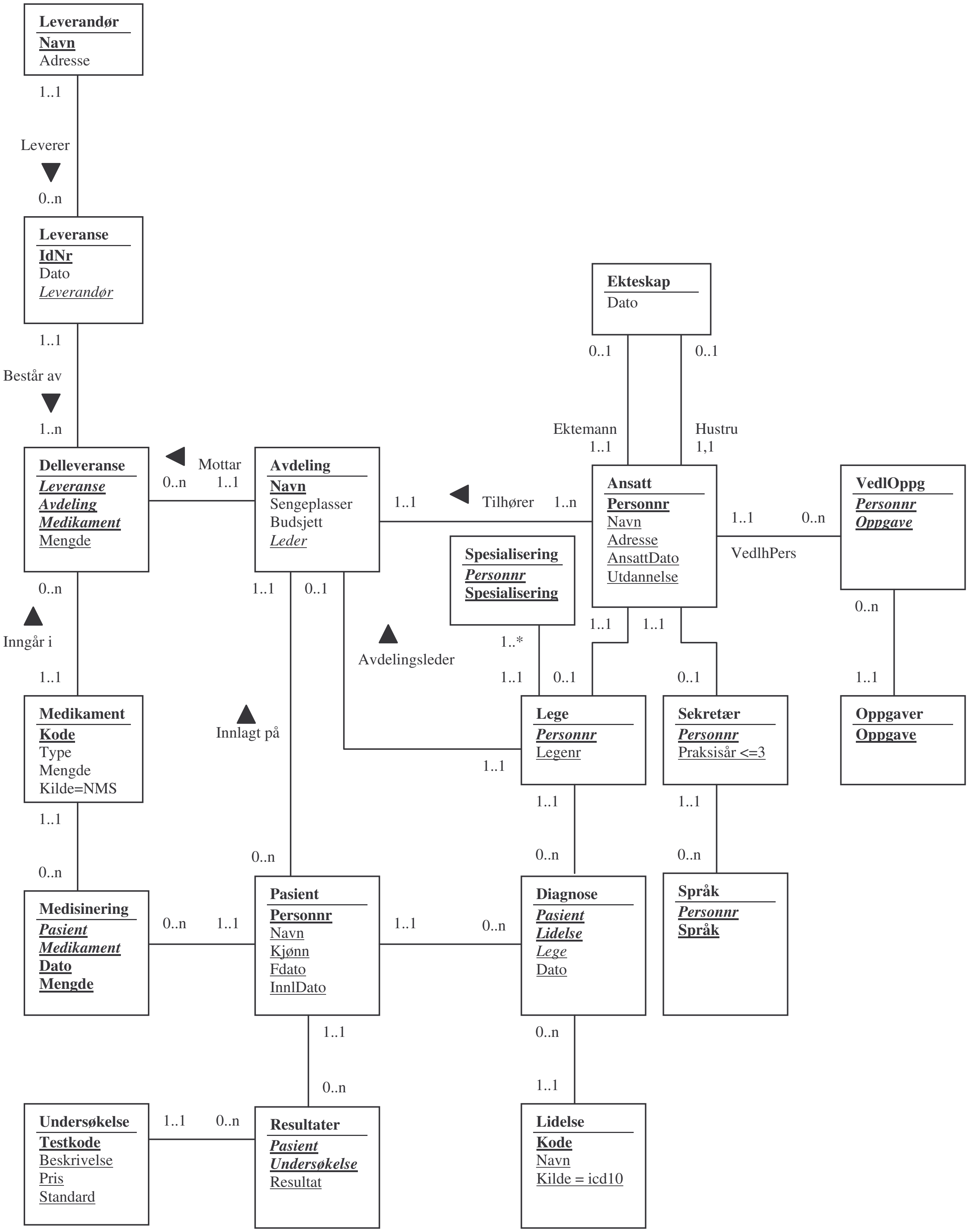
Lidelse (**Kode**, Navn, Kilde)

Diagnose (***Pasient, Lidelse***, Dato, *Lege*)

Undersøkelse (**Testkode**, Beskrivelse, Pris, Standard)

Resultater (**Pasient, Undersøkelse, Dato**, Resultat)

Skisse av den logiske modellen



Appendix E

Terminology database

A Terminology Database to Support Classification

TOR KRISTIAN BJELLAND
Institute for Information Science

University of Bergen
Postbox 7800, N-5020 Bergen, Norway
NORWAY
Tor.Bjelland@ifi.uib.no   http://www.ifi.uib.no/personer/ansatte/tor.html

Abstract: - Classification is generally held to be of fundamental importance to the analysis and design of software applications. But what exactly do we mean by classification? Based on a preliminary analysis of classification from conceptual modelling, archaeology, the cognitive sciences, and logic, I argue that classification has not received sufficient attention with respect to conceptual modelling, and that important distinctions between definition of terms, and description of objects have been overlooked. As a result, conditions for class membership are generally not specified. I assume that lack of membership conditions may impede communication, mediation, sharing, and re-use of of information, and seriously hamper schema integration and interoperability among applications. Consequently I suggest that classification should be introduced as a prerequisite to conceptual modelling. Hence, the main contribution of my research is to develop a theory of classification to guide the identification, definition and evaluation of the concepts and taxonomies that constitute the conceptual model.

Key-Words: - IS Models, Data Models, Classification, User/Analyst Differences, Information Analysis Techniques, IS Development Methods and Tools, Data Dictionary, Terminology

1 Introduction

Classification is generally recognized as a fundamental abstraction mechanism for conceptual modelling, and software engineering [4], [11]. This is clearly reflected in data modelling terminology, in which terms like classification, concept, class, superclass, subclass, IS_A relationship, generalization and specialization are frequently used. Classification is also considered to be the hardest part of analysis and design [4]. Yet, in spite of its importance, the discipline seems to lack a unified account of classification. As a result, the discipline is unable to provide simple answers to what classification means, how objects and classes are identified, and how class structures should be arranged and evaluated. As an example, consider the quote from Grady Booch, one of the leading figures among object-oriented methodologists:

> "Classification is the means whereby we order knowledge. In object-oriented design, recognizing the sameness among things allows us to expose the commonality within key abstractions and mechanisms, and eventually leads us to smaller and simpler architectures. Unfortunately, there is no golden path to classification. To the readers accustomed to finding cookbook answers, we unequivocally state that there are no simple recipes for identifying classes and objects. There is no such thing as a perfect class structure, nor the right set of objects...
> At a conference on software engineering, several developers were asked what rules they applied to identify classes and objects. Stroustrop, the designer of C++, responded: "It's a Holy Grail. There is no panacea". Gabriel, one of the designers of CLOS, stated, "That's a fundamental question for which there are no easy answer. I try things"." [4].

From the quote above, it seems reasonable to conclude that conceptual modelling, as a discipline, is in need of a coherent discourse on terms like classification and related notions, including concept, class, object and property. To be able to rigorously reason about model constructs, to provide answers to questions about modeling, and justifications for claims, such as the ones cited above, it becomes necessary to specify the domain of discourse, in a logically consistent and coherent manner, [15]. The need for a unified vocabulary has been articulated by several authors, as expressed in the following quote:

> "Snyder notes that; "…the groups involved with OO lack a shared understanding of the basic concepts and a common vocabulary for discussing them". Yourdon warns that: "…there is still enormous variation (and some contradictions as well) between the notation, strategies, and semantics of the various OOAD methodologies"… Discussing inheritance, Winkler notes that: "…this key-concept of OOP is interpreted quite differently by different groups of the software community"… Ling and Teo also recognize the lack of standards as one of the main inadequacies in OO data models." [16]

The motivation for this research is to advance the knowledge of classification, as it pertains to conceptual modelling. It is my conviction that a theory of classification will provide a common ground for rational discourses on key concepts like object and property, concept and class, definition and description, classification and inheritance.

The overall research approach to theory development is based on concept analysis, statement analysis and theory analysis of existing classification theories from related disciplines. Prototyping will be used for theory testing and revision. The theory will be implemented as a terminology meta-database, holding information about concepts, concept relationships, properties, applications and users.

2 Background

From a review of classification theories, it seems reasonable to distinguish between a cognitive, and a logical sense of classification. In the cognitive sense, classification is concerned with how people conceptualise the world, in the form of mental representations and operations. In the logical sense, classification is concerned with the definition of terms in order to concretise concepts. The main difference is that in the cognitive sense, concepts are subjective and private, while in the logical sense concepts are public, and hence, made inter-subjectively available by intensional definitions.

It appears that classification in the cognitive sense is the justification for classification in the logical sense. Research within the cognitive sciences has repeatedly demonstrated that concepts in general are subjective and vague, and liable to change, both between individuals and, over time, within the same individual.

It is exactly this vagueness, instability, and subjectivity of mental concepts that cognitive theories of classification attempt to explain, and the logical theory attempts to overcome.

How does this relate to conceptual data modelling? First of all, the two senses of classification may be viewed as the starting and the end points of the conceptual data modeling process. Kroenke [10] speaks of a database as a model of the user's mental models. Schlaer and Mellor [13] view conceptual modeling as a process in which separate and sometimes conflicting conceptual frameworks are brought together. Hirschheim and Klein [9] describe conceptual modeling as the fusion of horizons of meaning, given by the users' and developers' pre-understanding.

Thus, in order to arrive at an inter-subjectively shared and agreed upon representation of the application domain, the user's concepts must be concretised and reconciled into a common vocabulary.

Hence, conceptual data modelling may be viewed as a process whereby the users' and developer's knowledge of the application domain is given a uniform and explicit representation, in the form of a conceptual data model. This model, in turn, may be understood as a symbolic representation of the key concepts that make up the domain, along with the semantic and structural relationships that hold between them.

The fact that concepts are symbolically represented does not necessarily mean they are intensionally defined. On the contrary, as commented by Bergamashi and Sartory [2], the idea of intensional definitions are almost unheard of in the conceptual model tradition. Rather, emphasis has normally been placed on the descriptive properties of objects. The definition of a class in conceptual modelling is generally understood as the definition of its descriptive properties.

Since the definitional properties necessarily must be the same for all objects in a class, they are usually considered redundant, and hence excluded from the class definition. As a consequence, the membership criteria, which are supposed to settle whether an object belongs to a class, will at best remain as a commentary in a data dictionary, and hardly ever be noticed during the design and implementation of the application. However, the definitional properties play a critical role in any class-based application for several reasons:

First, definitional properties serve to specify a given domain of discourse in a logically consistent and coherent manner. The logical structures that result from classification, may guide the selection and specification of generalization hierarchies in the conceptual model.

Second, definitional properties are the only means to control that objects that enter a class really belong there. To check for membership criteria is a fundamental requirement that should be formalized and controlled by the application, not by the user. If users are unaware the membership conditions for a class, incorrect instances may be recorded. Hence, for class-based applications, membership conditions should be controlled by a related identification method for each class.

Third, problems related to homonymous and synonymous terms are easily confused with differences in descriptions. This is especially evident when attempts are made to integrate two or more separate applications. Since we usually employ descriptive properties in the definition of classes, we tend to assume that two classes are homonymous if they have the same class term, but differ in their descriptive properties. However, by emphasizing the distinction between definition of terms on one hand, and description of objects on the other, it becomes evident that the objects that fall under a definition, and hence, constitute a class, may be variously described for different purposes, or from different perspectives, while still being of the same kind.

In order to include definitional properties in conceptual modelling, I argue that classification is a prerequisite for conceptual modelling, and that the modelling process should be divided into two separate tasks: the classification process, which is concerned with the definition of concepts, and the modelling process, where the concepts are interrelated and further analysed with respect to descriptive properties.

The result from the classification process is a controlled vocabulary. A controlled vocabulary is defined as a restricted set of terms, in which the terms are normally selected for use within an organization, for a given purpose, in a specific subject field, and defined by intensional definitions. In terminology, definition by intension is the preferred method to provide abstract and stable definitions, and to disambiguate multireferential terms [14]. Hence, terms become organized in a hierarchical system, which are strictly based on genus-species relations only. Such an organization may be termed a classification system.

However, since most concepts that appear in a data model do not belong to the same genus, it will normally be the case that the controlled vocabulary contains several classification systems, ranging from simple systems of two concepts, the genus term and a single species term, to more complex systems that may span several levels of specialization. A classification system then, may be understood as a subset of a controlled vocabulary.

3 Related work

The need to consider the defining properties of concepts, not only the descriptive ones, is not entirely new, though, as commented by Bergamaschi and Sartory [2], the idea is almost unheard of in the conceptual model tradition.

Hakim and Garrett [8], suggest combining object-oriented modelling concepts with description logics, in order to overcome a number of limitations that follow directly from the inability of current object-oriented languages to define concepts by their necessary and sufficient conditions. Description logic is a kind of KR-language, which is divided into two separate languages: a terminological language to define concepts and relationships between concepts, and an assertional language, to create and manipulate individuals. The distinction between a terminological and an assertional language parallels our intuition that conceptual modelling should be similarly divided into definitional and descriptive parts.

Terminology is also a central issue in the current research on ontologies for knowledge-based systems. An ontology is considered a fundamental tool to support interoperability between knowledge systems, i.e., when knowledge sources are fused into a combined resource, like for instance a data warehouse, or when knowledge are to be shared among several knowledge-bases. Gamper, Nejdl, and Wolpers [5] explores the commonalities and differences between ontologies and terminologies. Wand, Storey and Weber [18], use ontology to analyse the meaning of common conceptual modelling constructs, and Guarino and Welty [7] presents a formal ontology of properties, in which important distinctions between membership conditions, identity conditions, object identifiers and primary keys are discussed.

Finally, my suggestion to explore the relevancy of various classification theories to conceptual modelling coincides with suggestions from [4], [12], and [17], who all introduce theories of classification from the cognitive sciences.

4 A terminology database

The fact that most modelling and programming languages lack any explicit mechanisms to define classes intensionally, makes it necessary to consider alternative ways to capture the defining properties. One strategy is to store metadata about existing application terminologies in a separate database for metadata management as depicted in fig. 1.

4.1 Purpose

The overall purpose of the metadata application is to record and retrieve information about existing application terminologies in order to facilitate the mediation, sharing, and re-use of terms. The meta database is not intended to control or restrict terminologies, but rather to guide the selection of terms during design of new applications and to provide different perspectives on term usage during integration of applications.

4.2 Design foundations

We briefly discuss in this section some non-trivial relationships between terms and term usage as depicted in fig. 1. A more detailed account of the fundamental assumptions and design foundations for the terminology database can be found in [3].

A language used for data modelling must be able to represent concepts of things, concepts of properties and concepts of relationships by means of terms and associated definitions. A term in this context is a designation consisting of one or more words representing a general concept of things, properties or relationships.

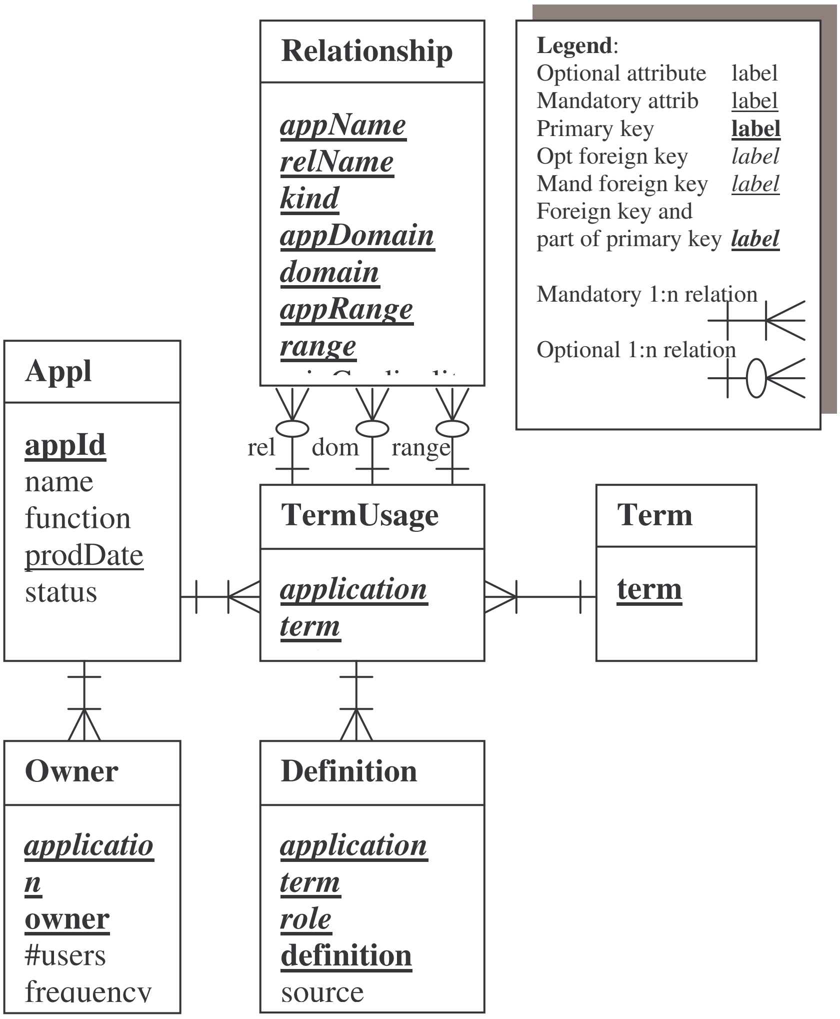


Fig. 1: A logical data model of the terminology data base.

A term *must* participate in at least one of the three roles depicted in fig. 1 between TermUsage and Relationship. Due to limitations in the notation that is used, it should be noted that this constraint is not properly expressed in the logical data model. Terms are used to name relationships, or the domain of a relationship, or the range of a relationship.

With respect to relationships, we distinguish between five kinds:

Identity relationships, which relate concepts of things to concepts of properties for identification purposes.
Descriptive relationships, which relate concepts of things to concepts of properties for descriptive purposes.
Partitive relationships, which relate superordinate concepts of things to subordinate concepts of things in a whole_part structure.
Taxonomic relationships, which relate superordinate concepts of things to subordinate concepts of things in a subsumption structure, where the intension of the subordinate concept includes the intension of the superordinate concept.
Associative relationships, which relate concepts of things when a thematic connection can be established by virtue of experience.

The different kinds of relationships may be considered as functions that map between a domain and a range. If the domain corresponds to a concept of things and the range corresponds to a concept of properties, the function is called an attribute. Hence, we may distinguish between identity attributes, used for identification purposes, and descriptive attributes used to describe concepts, or class members. If both the domain and the range correspond to concepts of things, then the relationship is either a partitive relationship, a taxonomic relationship or an associative relationship. A relationship is uniquely designated by the combination of the terms designating the relationship name and domain name.

Terms are further connected to the applications in which they occur, to reflect the application context of the terms. In an idealized world, where people would agree on concepts and terms, it would be unneccesary to include a reference to the application in which the terms appear. However, it is well recognized by the cognitive sciences that people tend to disagree both on concepts and on the use of terms. When people are asked to list all the relevant properties of a concept, the produce a tremendous amount of properties, including parts, compositions, superordinates, subordinates, origins, related objects, operations, actions, functions, beliefs, frequency and so forth [1]. Applications are further connected to application owners such as departments or organizational units to reflect the overall centrality of terms. Terms in applications that have several owners and many frequent users may be more central to the organization than terms that occur in a single user application. Finally the definition of each term is explicitly represented, along with an optional entry for authoritative sources, and an algorithm to support the identification of instances to which the definition applies.

Assuming the terminology database is implemented in a relational database management system, the immediate utility of the database can be demonstrated by a set of simple SQL statements. However, its real utility remains to be empirically tested on real applications.

Retrieval of class name and associated attributes.

```
Select appDomain, domain, appName, relName, kind,
minCardinality, maxCardinality From Relationship
Where (domain = ‘Paper’) AND (kind = ‘id’ OR kind = ‘desc’);
```

Retrieval of taxonomic relations:

```
Select appName, relName, appDomain, domain, appRange,
range From Relationship
Where (range = ‘Paper’) AND (kind = ‘Taxonomy’);
```

Retrieval of partitive relations:

```
Select appDomain, Domain, appName, appRel,
appRange, range From Relationship
Where (domain = ‘Session’) AND (kind = ‘Partitive’);
```

Retrieval of associative relations:

```
Select appDomain, domain, relName, appRange, range,
minCardinality, maxCardinality
From Relationship
Where (domain = ‘Person’) AND (kind = ‘Associative’);
```

Retrieval of information to detect different definitions of the same class term:

```
Select application, term, definition, source, algorithm
From Definition
Where term = ‘Paper’ AND role = ‘Class’
```

Retrieval of information to detect different descriptions of the same class:

```
Select appDomain, domain, appName, relName
From Relationship
Where domain = ‘Paper’ AND kind = ‘desc’ OR kind = ‘id’
```

5 Conclusion

Classification is generally held to be of fundamental importance to the analysis and design of software applications. However, as has been suggested in this paper, classification has not received sufficient attention with respect to conceptual modelling and analysis of software applications. Important distinctions between the definition of class terms, and the description of objects have not been noticed.

As a consequence, the membership criteria, which are supposed to settle whether an object is a member or not, will at best remain as a commentary in a data dictionary, and hardly ever be considered during the design and implementation of the application. Hence, it becomes possible that the responsibility for deciding whether something belongs to a class or not, is removed from the application, and, instead, imposed on the users. If users are unaware the membership conditions for the classes, incorrect objects may be recorded.

We argue that classification is a prerequisite for conceptual modelling, and that the modelling process should be divided into two separate tasks: classification followed by modelling.

One strategy to support the classification task is to store metadata about terminologies in a separate database for metadata management. This strategy has been explored to some extent from a theoretical perspective. However, it remains to be empirically tested. A different approach would be to include membership conditions and identification methods directly with the associated classes. This second approach is not considered in this paper, but suggests a direction for future research.

References:

[1] Barsalou, L.W. (1992): *Cognitive psychology : an overview for cognitive scientists*. L. Erlbaum, Hillsdale, N. J.

[2] Bergamaschi, S. & Sartori, C. (1992), On Taxonomic Reasoning in Conceptual Design, *ACM Transactions on Database Systems* Vol. 17, No. 3, pp. 385-422.

[3] Bjelland, Tor K. (2001), Classification: Basic Assumptions and Implications for Conceptual Data Modelling. To appear in *Proceedings of the 24th Information Systems Research Seminar in Scandinavia, Aug 11-14 2001*. Dept. of Information Science, University of Bergen, Norway.

[4] Booch, G. (1994), *Object-Oriented Analysis and Design With Applications*. Benjamin/Cummings Publishing Company, Inc. 2nd edition.

[5] Gamper, J., Nejdl, W. & Wolpers, M. (1999), Combining Ontologies and Terminologies, *Information Systems*. [online],[cited 25.08.00]. Available from Internet<http://www.kbs.uni-hannover.de/Arbeiten /Publicationen/1999/tke99>

[6] Grossmann, R. (1992), *The Existence of the World. An Introduction to Ontology*, Routledge, London and New York.

[7] Guarino, N. & Welty, C. (2000), A Formal Ontology of Properties, Proceedings of 12th Int. Conf. On Knowledge Engineering and Knowledge Management, Lecture Notes on Computer Science, Springer.

[8] Hakim, M.M. & Garrett, J.H.jr. (1997), An object-centered approach for modelling engineering design products: Combining description logic and object-oriented modelling, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, Vol 11, 187-198.

[9] Hirschheim, R. & Klein, H.K. (1995*): Information Systems Development and Data Modeling. Conceptual and Philosophical Foundations*. Cambridge University Press.

[10] Kroenke, D. (1998), *Database Processing. Fundamentals, Design, and Implementation*. 6th ed. Prentice-Hall.

[11] Mylopoulos, J. (1998), Information modelling in the time of the revolution, *Information Systems* Vol. 23, No.3/4, pp. 127-155.

[12] Parsons, J. (1996), An Information Model Based on Classification Theory, *Management Science*, 42(10) 1437-1453.

[13] Shlaer, S. & Mellor, S.J. (1988): *Object-oriented Systems Analysis. Modeling the World in Data*, Yourdon Press Computing Series, Englewood Cliffs, New Jersey.

[14] Svetonius, E. (1999): *The Intellectual Foundation of Information Organization*, MIT Press, Cambridge, Massachusetts.

[15] Sutcliffe, J.P. (1993): Concept, Class, and Category in theTradition of Aristotle, Van Mechelen et. al. (eds), *Categories and Concepts: Theoretical views and inductive data analysis.* Academic Press, London. MIT Press, Cambridge, Massachusetts.

[16] van Hillegersberg, J. & Kumar, K. (1999), Using Metamodeling to Integrate Object-Oriented Analysis, Design and Programming Concepts, *Information Systems* Vol. 24, No.3, pp. 113-129.

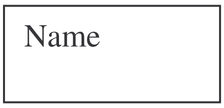
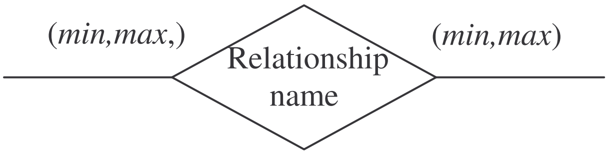
[17] Wand, Y., Monarchi, D.E., Parsons, J. & Woo, C.C. (1995), Theoretical foundations for conceptual modelling in information systems development, *Decision Support Systems* Vol. 15, pp. 285-304.

[18] Wand, Y., Storey, v.c. & Weber, R. (1999), An Ontological Analysis of the Relationship Construct in Conceptual Modeling, *ACM Transactions on Database Systems*, Vol. 24, No. 4, pp. 494-528.

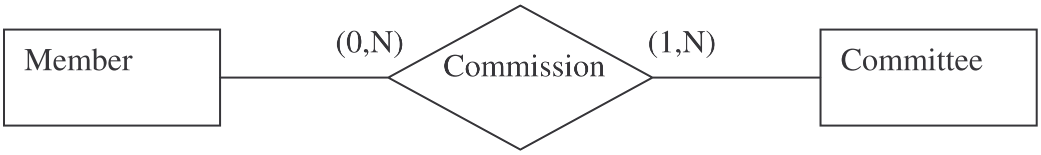
Appendix F

Alternative diagrammatic notations used in the thesis

Explanations to figure 2.7, 2.8, 4.18, and 4.19:

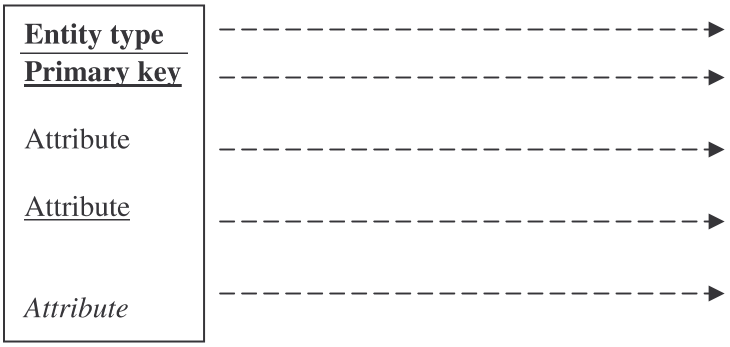
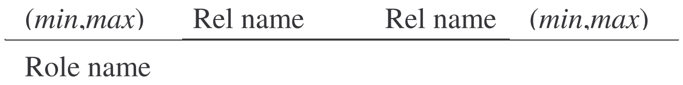
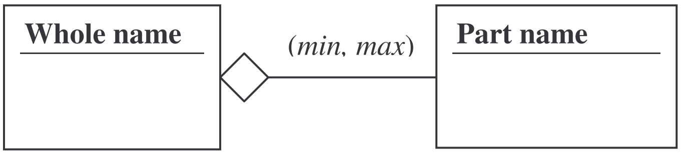
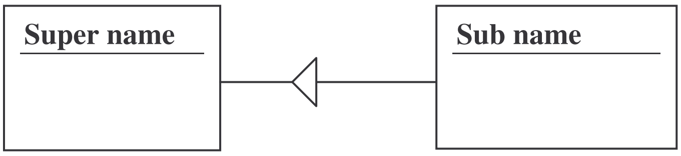
Symbols	**Descriptions**
	Entity type
	Relationship type
(min,max)	Cardinality constraints: *Min* = optionality, 0|1 *Max* = multiplicity, 1|n|*n*

Example:

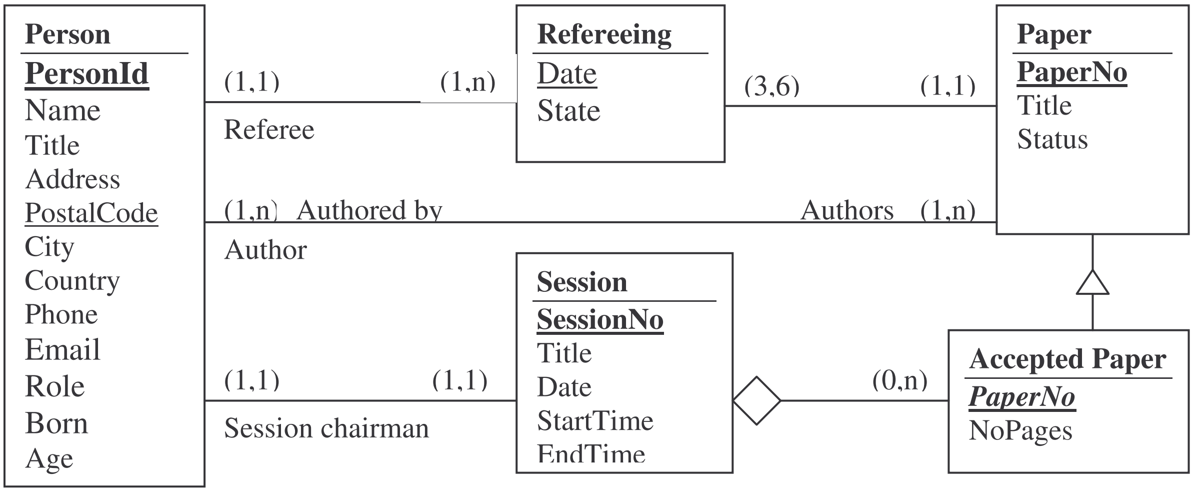


A commission is a relationship between Member and Committee. A member may be related to one committy. A committy must be related to many members.

Explanations to figure 4.20:

Symbols	Descriptions
 **Entity type** **Primary key** Attribute Attribute *Attribute*	Entity type name Primary key Attribute name (optional) Attribute name (mandatory) Foreign key (optional)
 (*min*,*max*) Rel name Rel name (*min*,*max*) Role name	Relationship with reversely positioned cardinality constraints. Relation names and role names are optional.
(min, max)	Cardinality constraints: *Min* = optionality, 0|1 *Max* = multiplicity, 1|n|*n*
	Partonomic relationship
	Generalization relationship

Example:



A person may either be a referee, an author, or a session chairman. This is reflected by the three role names, as well as by the role attribute in the person entity type. If a person is a referee, that person must review at least one, but possibly several papers, and a paper must be refereed by at least 3 and at most 6 referees. The relationship names have not been included due to space limitations.
A session may consist of accepted papers, and an accepted paper belongs to a subset of all papers that have been submitted for review.

References

Abrial, J. (1974) ‘Data Semantics’. In Klimbie and Koffeman (eds), *Data Base management*. North Holland.

Adams, W.Y. (1988) ‘Archaeological classification: theory versus practise’, *Antiquity* **61**, 40-56.

Adams, W.Y. and Adams, E.W. (1991) *Archaeological typology and practical reality: a dialectical approach to artifact classification and sorting*. Cambridge University Press.

Allen, R.B. (1997) ‘Mental Models and User Models’, in Helander, T.K., Landauer, P. and Prabhu, P. (eds): *Handbook of Human-Computer Interaction*. 2nd ed. Elsevier Science B.V.

Ambler, SW. (1995) *The Object primer. The Application developer’s Guide to Object-Orientation*. SIGS Books.

Artz, J. (1997) ‘A crash course on metaphysics for the database designer’, *Journal of Database Management*, **8**(4), 25-30.

Atzeni, P. (1999) *Database systems: concepts, languages and architectures*, McGraw-Hill. London.

Audi, R. (1995) *The Cambridge Dictionary of Philosophy*, Cambridge University Press.

Bailey, K.D. (1994) ‘Typologies and Taxonomies. An Introduction to Classification Techniques’, *Quantitative Applications In the Social Sciences*. Sage Publications, London.

Barnes, B. (1984) ‘The Conventional Component in Knowledge and Cognition’. In Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden*. SVT, UiB, Bergen 1996.

Barsalou, L.W. (1987) ‘The instability of graded structure: implications for the nature of concepts’. In Neisser, U. (Ed), *Concepts and conceptual development: Ecological and intellectual factors in categorization*, Cambridge University Press.

Barsalou, L.W. (1992) *Cognitive psychology : an overview for cognitive scientists*. L. Erlbaum, Hillsdale, N. J.

Barsalou, L.W. and Hale, C.R. (1993) ‘Components of Conceptual representation: From Feature Lists to Recursive Frames’, in Van Mechelen, et. al. (eds): *Categories and Concepts: Theoretical views and inductive data analysis.* Academic Press, London.

Batra, D, Hoffer, JA., and Bostrom, RP. (1990), ‘Comparing Representations with Relational and EER Models’, *Communications of the ACM*, February 1990, **33**(2), 126-139.

Batra, D. and Antony, R. (1994), ‘Effects of data model and task characteristics on designer performance: a laboratory study’, *Int. J. Human-Computer Studies*, **41**, 481-508.

Beck, C. and Jones, G.T. (1989) ‘Bias and Archaeological Classification’, *American Antiquity* **54**(2), 244-262.

Bell, D. and Morrey, I.. and Pugh, J. (1992) *Software Engineering. A Programming Approach*. 2nd ed. Prentice Hall.

Bergamaschi, S. and Sartori, C. (1992), ‘On Taxonomic Reasoning in Conceptual Design’, *ACM Transactions on Database Systems* **17**(3), 385-422.

Bock, H. H. (1994) ‘Classification and Clustering: Problems for the Future’, in Diday, E., et al (ed), *New Approaches in Classification and Data Analysis*. Studies in classification, data analysis, and knowledge organization. Springer-Verlag.

Boman, M., Bubenko, J.A. Jr., Johannesson, P., and Wangler, B. (1997) *Conceptual Modelling*. Prentice Hall.

Booch, G. (1991), *Object-Oriented Analysis and Design With Applications*. Benjamin/Cummings Publishing Company, Inc.

Booch, G., Rumbaugh, J. and Jacobson, J. 1999. *The Unified Modeling Language User Guide*. Addison-Wesley.

Borgida, A. and Myloupolos, J. and Wong, H.K.T. (1986) ‘Generalization/Specialization as a Basis for Software Specification’, in Brodie et.al. (eds): *On Conceptual Modelling*. Springer-Verlag, Virginia.

Bowker, G. C. and Leigh Star, S. (1998) ‘How things (actor-net)work: Classification, magic and the ubiquity of standards’. Spring. [online],[cited 22.01.98]. Availiable from Internet <URL: http://alexia.lis.uiuc.edu/¨bowker/actnet.html>

Bowker, L. and Lethbridge, T.C. (1994) ‘Terminology and Faceted Classification: Applications Using CODE4’. *Advances in Knowledge Organization* **4**, 200-207.

Brodie, ML. (1986) ‘On the Development of Data Models’. In Brodie, M.L. and Mylopoulos, J. and Schmidt, J.W. (Eds), *On Conceptual Modelling*, Springer-Verlag.

Brodie, M. and Mylopoulos, J. and Schmidt, J.W. (1986) *On Conceptual Modelling*. Springer-Verlag, Virginia.

Brodie, ML, and Ridjanovic, D. (1986) ‘On The Design and Specification of Database Transactions’. In Brodie, M.L. and Mylopoulos, J. and Schmidt, J.W. 1986. *On Conceptual Modelling*. Springer-Verlag.

Bubenko, J. Jr. (1977) ‘IAM: An Inferential Abstract Modeling Approach to the Design of Conceptual Schema’, in Smith, D.C.P. (Ed), *Proceedings of the 1977 ACM Sigmod Conference on Management of Data*, Aug. 3-5, Canada.

Bubenko, J. Jr. (1980) ‘Information Modelling in the Context of System Development’, in Lavington, SH (ed), *Information Processing*, North-Holland.

Bubenko, J. Jr. and Lindencrona, E. (1984) *Konceptuell modellering – Informationsanalys*. Studentlitteratur.

Bunge, M. (1977) *Treatise on Basic Philosophy*. **3**. Ontology 1: The furniture of the world, D. Reidel Publishing Company

Bunge, M. (1983) *Treatise on Basic Philosophy*. **5**. Epistemology and Methodology I: Exploring the World, D. Reidel Publishing Company.

Cattell, R.G.G. and Barry, D.K. (2000) *The Object Data Standard ODMG 3.0*. Morgan Kauffmann.

Carmichael, A.(1994) *Object Development Methods*. Sigs Books.

Carmichael, A. (1997) *Developing Business Objects*. Sigs Books.

Chen, PP. (1976). ‘The Entity-Relationship Model – Toward a Unified View of Data’, *ACM Transactions on Database Systems*, **1**(1), March 1976, 9-36.

Chen, PC. (1977): *The Entity-Relationship Approach top Logical Database design. The Original Work of Dr. Chen*. QED Information Sciences Inc.

Coad, P. and Nicola, J. (1993): Object-Oriented Programming. Yourdon Press.

Coad, P. and Yourdon, E. (1991): *Object-Oriented Analysis*. Yourdon Press, 2nd ed.

Codd, E. (1979): ‘Extending the Database Relational Model to Capture More Meaning’, in *Transactions on Database Systems*, **4**(4), December 1979.

Connolly T., Begg C., and Strachan A. (2002) *Database Systems. A Practical Approach to Design, Implementation and Management*. Addison Wesley, 3rd ed.

Copi, I.M. and Cohen, C. (1998) *Introduction to Logic*. 10th ed. Prentice-Hall, N.J.

Date, C.J. (2000) *An Introduction to Database Systems*. Addison Wesley. 7th ed.

Douglas, M. and Hull, S. (1992) *How classification works. Nelson Goodman among the social sciences*, Edinburgh University Press, Edinburgh.

Dunn, C.L. and Grabski, S.V. (2001) ‘Syntactic and Semantic Understanding of Conceptual Data Models’, *2001-Twenty Second International Conference on Information Systems*, New Orleans.

Dunnell, R.C. (1994) *Systematics in Prehistory*. Spring. [online],[cited 08.06.99]. Availiable from Internet <URL: http://weber.u.washington.edu/¨antro/BOOK/book.html>

Edmond, D. (1992) *Information Modeling. Specification and Implementation*. Prentice-Hall, Australia.

Elmagarmid, A., Rusinkiewicz, M. and Sheth, A (1999) *Management of Heterogeneous and Autonomous Database Systems*. Morgan Kaufmann Publishers, San Francisco, California.

Elmasri, R. and Navathe, S.B. (2000) *Fundamentals of database Systems*. 3rd ed. Benjamin/Cummings Publishing Company.

Embley, D.W., Kurtz, B.D. and Woodfield, S.N. (1992) *Object-Oriented Systems Analysis. A Model-Driven Approach*. Yourdon Press Computing Series.

Embley, D.W. (1998) *Object Database development. Concepts and Principles*. Addison-Wesley.

Eriksson, H.E. and Penker, M. (2000) *Business Modeling with UML. Business Patterns at Work*. John Wiley and Sons, Inc.

Everest, G.C. (1986) *Database management. Objectives, System Functions and Administration*. McGraw-Hill, Singapore.

Finkelstein, C (1990) *An Introduction to Information Engineering. From Strategic Planning to Information Systems*, Addison Wesley.

Gamper, J., Nejdl, W. and Wolpers, M. (1999), 'Combining Ontologies and Terminologies in Information Systems'. [online],[cited 25.08.00]. Available from Internet <http://www.kbs.uni-hannover.de/Arbeiten/Publicationen/1999/tke99>

Gatewood, J.B. (2000) 'Ignorance, Knowledge, and Dummy Categories: Social and Cognitive aspects of Expertise'. [online],[cited 13.04.00]. Availiable from Internet <URL: http://www.lehigh.edu/~jbg1/liquors.htm>

Gilje, N. (1996) 'Anomalier i moderne vitenskapsfilosofi', in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden.* SVT, UiB, Bergen 1996.

Gilje, N. and Grimen, H. (1996) 'Hermeneutikk, mening og forståelse', in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden.* SVT, UiB, Bergen 1996.

Green, R. (1992) 'Classificatory structures in cognitive science', in Williamson and Hudon (eds), *Classification Research for Knowledge Representation and Organization*. Elsevier Science Publishers B.V.

Grossmann, R. (1992), *The Existence of the World. An Introduction to Ontology*, Routledge, London and New York.

Gruber, T.R. (1995): Toward principles for the design of ontologies used for knowledge sharing. *International Journal Human-Computer Studies*, **43**, 907-928.

Guarino, N. and Welty, C. (2000), 'A Formal Ontology of Properties', in *Proceedings of 12th Int. Conf. On Knowledge Engineering and Knowledge Management*, Lecture Notes on Computer Sciernce, Springer.

Guba, E.G. (1990) *The Paradigm dialog*. Sage Publications, Newbury Park, California.

Guba, E.G. and Lincoln, Y.S. (1998) ‘Competing Paradigms in Qualitative Research’, in Denzin and Lincoln (Eds): *The Landscape of qualitative research : theories and issues.* Sage, Thousand Oaks, Calif.

Hahn, U. and Chater, N. (1997): ‘Concepts and Similarity’ in Lamberts, K. and Shanks, D. (eds): *Knowledge, Concepts and Categories*. Psychology Press. Birmingham.

Hakim, M.M. and Garrett, J.H.jr. (1997) ‘An object-centered approach for modelling engineering design products: Combining description logic and object-oriented modelling’, *Artificial Intelligence for Engineering Design, Analysis and Manufacturing*, **11**, 187-198.

Halpin, T. (2004) ‘Subtyping: Conceptual and logical issues’. Downloaded from: <URL:http://www.orm.net/pdf/Subtype.pdf >

Hammer, M. and McLeod, D. (1981) ‘Database Description with SDM: A Semantic Data Model’. *Transactions on Database Systems*‘, **6**(4), September 1981.

Hampton, J. (1993) ‘Prototype Models of Concept Representation’, in Van Mechelen, et. al. (eds): *Categories and Concepts*: *Theoretical views and inductive data analysis*. Academic Press, London.

Hampton, J. and Dubois, D. (1993) ‘Psychological Models of Concepts: Introduction’, in Van Mechelen, et. Al. (eds): *Categories and Concepts: Theoretical views and inductive data analysis*. Academic Press, London.

Hanson, N.R. (1969) ‘Theory-laden Language’, in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden*. SVT, UiB, Bergen 1996.

Hempel, C.G. (1994) ‘Fundamentals of Taxonomy’. Reprinted in Sadler, Wiggens and Schwartz (eds): *Philosophical Perspectives on Psychiatric Diagnostic classification*. John Hopkins University Press, Baltimore.

Hirschheim, R. and Klein, K.K. (1995). *Information Systems development and Data Modeling. Conceptual and Philosophical Foundations*. Cambridge University Press.

Hoffer, J.A., George, J.F. and Valacich, J.S. (2002) *Modern Systems Analysis and Design*. Prentice Hall.

Hunter, E.J. (1988) *Classification Made Simple*. Gower, Aldershot.

ISO 704:2000(E) (2000) *Terminology work – Principles and methods*.

Jacob, Elin. (1994) ‘Classification and Crossdisiplinary Communication: Breaching the Boundaries Imposed by Classificatory Structure’, in Albrechtsen and Oernager (eds), *Knowledge Organization and Quality Management*. Advances in Knowledge Organization, **4**. Index Verlag, Frankfurt/Main.

Jacobsen, I., Christerson, M., Jonsson, P. and Øvergaard, G. 1994: *Object-Oriented Software Engineering. A Use-Case Driven Approach*. Addison Wesley.

Jacobsen, I., Booch, G. and Rumbaugh, J. (1999) *The Unified Software Development Process*. Addison Wesley.

Johnston, I. (2001): 'Some Non-Scientific Observations on the Importance of Darwin'. [online],[cited 12.01.2001]. Availiable from Internet <URL: http://www.mala.bc.ca/~johnstoi/essays/darwin.htm>

Kangassalo, H. (1992) 'On The Concept of Concept for Conceptual Modelling and Concept Detection', in Oshuga, S. et.al., (eds): *Information Modelling and Knowledge Bases III*, IOS Press.

Kent, W. (1978): *Data and Reality*. Basic Assumptions in Data Processing Reconsidered, North-Holland.

Kim, Y.G. and March, S.T. (1995) 'Comparing Data Modeling Formalisms', *Communications of the ACM*, June 1995, **38**(6), 103-115.

Klein, H.K. and Hirschhreim, R.A. (1987) 'A Comparative Framework of data Modelling Paradigms and Approaches'. The Computer Journal, **30**(1), 8-15.

Kroenke, D.M. (2002) *Database Processing. Fundamentals, Design, and Implementation*. 8th ed. Prentice-Hall, N.J.

Kuhn, T.S. (1974) 'Second Thoughts on Paradigms', in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden*. SVT, UiB, Bergen 1996.

Kuhn, T.S. (1991) 'The natural and the human sciences', in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden*. SVT, UiB, Bergen 1996.

Lakoff, G. (1987*) Women, Fire and Dangerous Things. What Categories Reveal about the Mind*. The University of Chicago Press, Chicago.

Lamberts, K. and Shanks, D. (1997) *Knowledge, Concepts and Categories*. Psychology Press. Birmingham.

Lane, R. (2002) 'PHIL 3120, American Philosophy. Lecture Notes'. Fall [online],[cited 01.12.2003] Available from Internet. <URL:http://www.westga.edu/~rlane/american/>

Langer, S. (1967) *An introduction to symbolic logic*. 3rd edn. New York.

Langridge, D.W. (1992) *Classification: Its kinds, elements, systems and applications*. Bowker Saur.

Lewis, P., Berstein, A. and Kifer, M. (2002) *Databases and Transaction Processing. An Application-Oriented Approach*. Addison Wesley.

Leigh Star, S. (1998): ‘Grounded classification: Grounded Theory and Faceted Classification’. Autumn. [online],[cited 25.05.98]. Availiable from Internet <URL: http://alexia.lis.uiuc.edu/¨star/gt.html >

Mac Gregor, R. (1991): ‘The Evolving Technology of Classification-Based Knowledge Representation Systems’, in Sowa, J.F., (ed), *Principles of Semantic Networks*. Morgan Kaufmann Publishers.

Malpas, J. (1987) *Prolog: A relational Language and its implementation*. Prentice Hall.

Malt, B.C. (1995) ‘Category Coherence in Cross-Cultural Perspective’. *Cognitive Psychology*. **29**, 85-148.

March, S.T. and Smith, G.F. 1995 ‘Design and Natural Science Research on Information Technology’. *Decision Support Systems*, **15**(4), Des 1995, 251-266

March, S.T. (2000) ‘Reflections on Computer Science and Information Systems Research’, in Laender, H.F., Liddle, S.W. and Storey, V. (eds): **Conceptual Modeling – ER 2000**. 19th International Conference on Conceptual Modeling. Springer.

Martin, J. and Odell, J.J. (1992) *Object-Oriented Analysis and Design. A Foundation*. Prentice Hall.

Martin, J. and Odell, J.J. (1996) *Object-Oriented Methods. Pragmatic Considerations*. Prentice Hall.

Martin, J. and Odell, J.J. (1998) *Object-Oriented Methods. A Foundation. UML edition*. Prentice Hall.

McCauley, R.N. (1987) ‘The Role of Theories in a theory of concepts’, in Neisser, U (ed): *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge University Press.

Medin, D.L., Wattenmaker, W.D. and Hampson, S.E. (1987) ‘Family Resemblance, Conceptual Cohesiveness, and Category Construction’. *Cognitive Psychology*, **19**, 242-279.

Medin, D.L. (1989) ‘Concepts and Conceptual Structure’. *American Psychologist*, **44**(12), 1469-1481, December 1989.

Medin, D.L. and Ross, B.H. (1997) *Cognitive Psychology*. Harcourt Brace College Publishers.

Medin, D.L, Lynch, E.B., and Solomon, K.O. (2000) ‘Are There Kinds of Concepts’? *Annu. Rev. Psychol*, **51**, 121-147.

Mellor, DH and Oliver, A. (1997) *Properties*. Oxford Readings in Philosophy. Oxford University Press.

Mineau, G., Stumme, G. and Wille, R. (1999) *Conceptual Structures Represented by Conceptual Graphs and Formal Concept Analysis*. Preprint No. 2034, Technische Universitet Darmstadt.

Moody, DL. and Shanks, GG. (1998) 'What Makes a Good Data Model? A Framework for Evaluating and Improving the Quality of Entity Relationship Models', *The Australian Computer Journal*, **30**(3), August 1998.
Murphy, G.L. and Medin, D.L. (1985) 'The Role of Theories in Conceptual Coherence'. *Psychological Review*, **92**(3), July 1985, 289-316.

Murphy, G.L. (1993) 'Theories and Concept Formation', in Van Mechelen, et. al. (eds): *Categories and Concepts: Theoretical views and inductive data analysis*. Academic Press, London.

Mylopoulos, J. (1998), 'Information modelling in the time of the revolution', *Information Systems* **23**(3/4), 127-155.

Neisser, U. (1987). *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge University Press.

Norrie, M. (2000): 'Advances in Object-Oriented Data Modeling', in Papazoglu, M.P., Spaccapiera, S, and Zahir, T. (eds), *Advances in Object-Oriented Data Modeling*. MIT Press.

Odell, J. and Ramackers, G. (1997): *Toward a Formalization of OO Analysis*. Downloaded from http://www.quoininc/JOarticle9707.html.

Odell, JJ. (1998) *Advanced Object-Oriented Analysis and Design Using UML*. SIGS Reference Library. Cambridge University Press.

Ogden, C.K. and Richards, I.A. (1972) *The meaning of meaning : a study of the influence of language*.10th ed. Routledge and Kegan Paul, London.

Olle, W. (1988) 'System Design Specifications for a Conference Organization System', in *Proceedings of the IFIP WG 8.1 Working Conference on Computerized Assistance during the Information Systems Life Cycle*, CRIS 88, Egham England, North-Holland.

Orlikowsky, W.J. (1995) 'Categories: Concept, Content, and Context'. *Computer Supported Cooperative Work (CSCW)*, **3**(1), 73-78.

Page-Jones, M. (2000) *Fundamentals of Object-Oriented Design in UML*. Addison Wesley.

Parsons, J. (1996), 'An Information Model Based on Classification Theory'. *Management Science*, 42(10) 1437-1453.

Parsons, J. and Wand Y. (1997a), Choosing Classes In Conceptual Modeling, *Communications of the ACM*, June 1997, **40**(6), 63-69.

Parsons, J. and Wand Y. (1997b), Using Objects for Systems Analysis, *Communications of the ACM*, December 1997, **40**(12), 104-110.

Parsons, J. and Wand Y. (2000), Emancipating Instances from the Tyranny of Classes in Information Modeling, *ACM Transactions on database Systems*, **25**(2), June 2000, 228-268.

Patridge, C. 1996: *Business Objects. Re-engineering for re-use*. Butterworth-Heinemann.

Polit, DF. and Hungler, BP. (1999): *Nursing Research. Principles and Methods*. 6th ed. Lippincott.

Picht, H. and Draskau, J. (1985): *Terminology: An Introduction*. The University of Surrey, Dept. of Linguistics and International Studies. Guilford, Surrey, GU2 5XH, England.

Quine, W.V.O. (1977) 'Natural Kinds', in S.P. Schwartz (Ed): *Naming, Necessity and Natural Kinds*. Cornell University Press, Ithaca, NY.

Rahm, E. (2001) 'A survey of approaches to automatic schema matching'. *The VLDB Journal*, **10**, 334-350.

Ram, S. and Ramesh, V. (1999) 'Schema Integration: Past, Present, and Future', in Elmagarmid, A. et. al. (eds): *Management of Heterogeneous and Autonomous Database Systems*. Morgan Kaufmann Publishers.

Resnick, L.B., Levine, J.M. and Teasley, S.D. (1991). *Perspectives of Socially Shared Cognition*. American Psychological Association. Washington DC.

Robinson, K. and Berrisford, G. (1994) *Object-Oriented SSADM*. Prentice Hall.

Rodgers, B.L. (2000) 'Concept Analysis: An Evolutionary View', in Rodgers, B.L. and Knafl, K.A. (eds): *Concept Development in Nursing. Foundations, Techniques, and Applications*. 2nd ed. W.B. Saunders Company.

Rodgers, B.L. and Knafl, K.A. (2000) *Concept Development in Nursing. Foundations, Techniques, and Applications*. 2nd ed. W.B. Saunders Company.

Rosch, E. and Lloyd, B.B. (1978): *Cognition and categorization*. Erlbaum, Hillsdale N.J.

Rosch, E. (1978) 'Principles of Categorization' in Rosch, E. and Lloyd, B.B. (eds*) Cognition and Categorization*. Lawrence Erlbaum Associates, New Jersey.

Rumbaugh, J., Jacobson, I. and Booch, G. (1999) *The Unified Modeling Language reference manual.* Addison Wesley.

Ryle, G. (1951) *Thinking and Language, Proceedings of the Aristotelian Society* (Supplementary Series), **25**, 65-82.

Sarantakos, S. (1998) *Social Research*. 2nd ed. MacMillan Press Ltd.

Schwandt, T.A. (1998) 'Constructivist, Interpretivist Approaches to Human Inquiry', in Denzin and Lincoln (Eds): *The Landscape of qualitative research : theories and issues*. Sage, Thousand Oaks, Calif.

Schøn, D.A. (1967) *Technology and Change*. Delacorte Press. New York.

Shapere, D. (1981) ‘Meaning and Scientific change’, in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden*. SVT, UiB, Bergen 1996.

Shaw, M.L.G. and Gaines, B.R. (1998) *Comparing Conceptual Structures: Consensus, Conflict, Correspondence and Contrast*. Fall, [online],[cited 27.10.98]. Availiable from Internet <URL: http://ksi.cpsc.ucalgary.ca/articles/KBS/COCO/>

Shlaer, S., and Mellor, S.J. (1988) *Object-Oriented Systems Analysis. Modelling the World in Data*. Yourdon Press.

Shipman III, F.M. and Marshall C.C. (1999) ‘Formality Considered Harmful: Experiences, Emerging Themes, and Directions on the Use of Formal Representations in Interactive Systems’, *Computer Supported Cooperative Work,* **8**, 333-352. Kluver Academic Publishers.

Shoval, P. and Shiran, S. (1997) ‘Entity-Relationship and object-oriented data modelling – an experimental comparison of design quality’, *Data and Knowledge Engineering* **21**, 1997, 297-315.

Siau, K., Wand, Y. and Benbasat, I. (1995) ‘A Psychological Study on the Use of Relationship Concept – Some Preliminary Findings’, *7th International Conference on Advanced Information Systems*, **932**(932), 341-354.

Siau, K., Wand, Y. and Benbasat, I. (1997) ‘The Relative Importance of Structural Constraints and Surface Semantics in Information Modeling’, *Information Systems*, **22**(2/3), 155-170.

Smith, J.M. and Smith, D.C.P. (1977) Database Abstractions: Aggregation and Generalization’, in *ACM Transactions on Database Systems*, **2**(2), 105-133, June 1977.

Smith, E.E. and Medin, D.L. (1981) *Categories and Concepts*. Harvard University Press. Cambridge.

Sokal, R, R. (1974): ‘Classification: Purposes, Principles, Progress, prospects’. *Science*, **185**(4157), September 1974, 1115-1123.

Solomon, K.O, Medin, D.L. and Lynch, E. (1999) ‘Concepts do more than categorize.’ *Trends in Cognitive Sciences*, **3**(3), 99-105, March 1999.

Sowa, J.F. (1984) *Conceptual Structures: Information Processing in Mind and Machine*. Addison-Wesley.

Sowa, J.F. (2000) *Knowledge Representation. Logical, Philosophical and Computational Foundations*. Brooks/Cole.

Star, S.L. and Bowker, G.C. (2000) *Sorting Things Out. Classification and its Consequences*. MIT-Press. Cambridge, Massachusetts, London, England.

Starr, P. (1992) ‘Social categories and Claims in the Liberal State’, in Douglas and Hull (Eds): *How classification works. Nelson Goodman among the social sciences*. Edinburgh University Press, Edinburgh.

Stonebraker, M. and Hellerstein, JM (1998) *Readings in Database Systems*. 3rd ed. Morgan Kaufmann.

Storms, G. and De Boeck P. (1997) ‘Formal Models for Intra-categorical Structure that can be Used for data Analysis’, in Lamberts, Koen and Shanks, (eds): *Knowledge Concepts and Categories,* Psychology Press.

Suppe, F. (1977) *The Structure of scientific theories. Edited with a critical introduction and an afterword by Frederick Suppe.* - 2nd ed. University of Illinos Press, Urbana.

Suppe, F (1989) *The semantic conception of theories and scientific realism.* University of Illinois Press, Chicago.

Sutcliffe, J.P. (1993) ‘Concept, Class, and Category in the Tradition of Aristotle’, in Van Mechelen et. al. (eds): *Categories and Concepts: Theoretical views and inductive data analysis*. Academic Press, London.

Sutcliffe, J.P. (1994) ‘On the logical necessity and priority of a monothetic conception of class, and on the consequent inadequacy of polythetic accounts of category and categorization’, in Diday, E., et al (eds): *New Approaches in Classification and Data Analysis. Studies in classification, data analysis, and knowledge organization*. Springer-Verlag.

Svetonius, E. (1999) *The Intellectual Foundation of Information Organization*. MIT Press, Cambridge, Massachusetts.

Sølvberg, A. and Kung, D.C. (1998) *Information Systems Engineering. An Introduction*. Springer.

Taylor, C. (1980) ‘Understanding in Human Science’, in Gilje and Grimen (eds): *Kompendium i almen vitskapsteori for Dr.Polit. og Dr.Art.-graden.* SVT, UiB, Bergen 1996.

Thagard, P. (1992) *Conceptual Revolutions*. Princeton University Press.

Tsichritzis, D.C. and Lochovsky, F.H. (1982) *Data Models*. Prentice Hall.

Ullman, J. and Widon, J. (1997) *A First Course in Database Systems*. Prentice Hall.

Van Hillegersberg, J. and Kumar, K. (1999) ‘Using Metamodeling to Integrate Object-Oriented Analysis, Design and Programming Concepts’. *Information Systems* **24**(3), 113-129.

Van Mechelen, I., De Boeck, P., Theuns, P. and Degreff, E. (1993) ‘Categories and Concepts: Theoretical Views and Inductive Data Analysis’ in Van Mechelen, I.. and Hampton, J., and Michalski, R.S., and Theuns, P (eds): *Categories and Concepts: Theoretical views and inductive data analysis*. Academic Press, London.

Van Mechelen, I., and Hampton, J., and Michalski, R.S., and Theuns, P. (1993) *Categories and Concepts: Theoretical views and inductive data analysis*. Academic Press, London.

Vickery, B.C. (1960) *Faceted Classification : A guide to construction and use of special schemes*. Aslib, London.

Vickery, B.C. (1975) *Classification and Indexing in Science*. Butterworths, London.

Wand,Y., Monarchi, D.E., Parsons, J. and Woo, C.C. (1995) 'Theoretical foundations for conceptual modelling in information systems development', *Decision Support Systems* **15**, 285-304.

Wand, Y., Storey, V.C. and Weber, R. (1999) 'An Ontological Analysis of the Relationship Construct in Conceptual Modeling', *ACM Transactions on Database Systems*, **24**(4), 494-528.

Waterson, A.and Preece, A. (1999) Verifying ontological commitment in knowledge-based systems, *Knowledge-Based Systems* **12**, 45-54.

Whittaker JC, Caulkins D, Kamp KA. (1998): 'Evaluating consistency in typology and classification'. *Journal Of Archaeological Method and Theory*, **5**(2), 129-164, June 1998

Wittgenstein, L. (1953) *Philosophical Investigations*. Macmillan. New York.

Woods, WA. (1991): 'Understanding subsumption and taxonomy', in Sowa, J.F. (ed): *Principles of Semantic Networks*. Morgan Kaufmann Publishers.