2002

Understanding Categorization: An Experientialist Perspective

Jennifer Sampson
University of Melbourne

UNDERSTANDING CATEGORISATION: AN EXPERIENTIALIST PERSPECTIVE

Jennifer Sampson

Department of Information Systems
University of Melbourne
Melbourne, Australia

ABSTRACT

A considerable amount of research has been undertaken to establish how end users interpret, understand and verify data models often with the motivation of improving understanding, accuracy and completeness. However, there is relatively little research examining how and why humans make categorisation decisions when developing information systems. The research question addressed in this paper is: what is it about the human mind that allows it to categorise concepts in a particular way for conceptual data modelling? From a cognitive psychology perspective Lakoff's (1987) image schema structures provide answers to such questions. I argue that acknowledging such structures is the first step in defining an experientialist strategy for data modelling. In addition a case example is used to illustrate the problems with utilising the objectivist approach to categorisation.

1. INTRODUCTION

Categorisation is the main way that we come to understand experience. Conceptual data modelling as a categorisation activity is essentially a matter of both human experience (perception, motor activity and culture) and imagination (metaphor, metonymy, and mental imagery). One question of interest is how do we do this? More specifically the research question addressed in this paper is what is it about the human mind that allows it to categorise concepts in a particular way for conceptual data modelling? I argue that concepts are not internal representations of external reality, because there is no corresponding reality or categories "out there" to be mirrored. This is in contrast to the objectivist paradigm which claims that concepts provide internal representations of external reality. In sections two and three of this paper I will briefly describe why the classical approach to categorisation is inappropriate for conceptual data modelling. Section three presents the experientialist strategy, an alternate approach for establishing a theoretical foundation for conceptual data modelling. An initial application of this approach is presented in section four. Finally suggestions for future research to empirically test the efficacy of the strategy for conceptual data modelling are discussed.

2. PREVIOUS RESEARCH

Ontology can be described as 'what there is', and relies on the use of specific terms to construct a description of reality. The use of ontology as a theoretical basis for data modelling is inappropriate because we are not modelling reality when developing information systems, but the way information about reality is processed, by people (Kent, 1978). Despite this, research has been undertaken to determine a theoretical foundation for data modelling based on philosophical ontologies [Wand and Weber 1993, Wand Storey and Weber 1999, Wand and Wang 1996, Weber 1997]. In particular, Wand *et al* (1999) claim that the use of Bunge's ontology is useful for enhancing the understanding and meaning of conceptual models. However, little empirical research has been undertaken to determine whether humans interpret ontologically based representations better than a representation which does not conform to an 'acceptable' ontology. In essence they have not proposed why an ontologically sound model might be more understandable than one which is not.

The traditional classical view of categorisation claims that a category can be described by a set of defining attributes. The classical view is based on an objectivist philosophy implying that boundaries of what is, and is not a category member, is clear cut. Furthermore it is an attempt to model reality independent of human categorisation, however the constructs used to create the model are only a product of the mind. The main problem with the classical view is how the defining attributes are decided upon, as almost all non-trivial categories have fuzzy boundaries. As Kent (1978) claims "it is often a matter of choice whether a piece of information is to be treated as a category, an attribute, or a relationship".

The basic metaphysical belief of objectivism is that categories exist in objective reality. However, there are no categories of the right kind objectively "out there" and there are no words (or symbols) which objectively match up with categories in the world. Bunge's ontology based on the objectivist paradigm is inappropriate as a theoretical basis for conceptual data modelling because it fails to incorporate bodily experience and imaginative processes. The objectivist view implies that the symbols of natural language have meaning because they can be explicitly related to objective categories. This account of meaning fails to mention human beings – "It does not depend in any way on the nature of the thinking and communicating organisms, or on the nature of their experience" (Lakoff, 1987).

In addition problems with utilising the objectivist paradigm for data modelling are that often the need for meaningfulness is lost on the data modeller. The reason may be that when developing a data model, the data modeller does so with her/his own understanding of what the entities, attributes and business rules are supposed to mean. For this reason, the model does not seem meaningless to the data modeller because the entities and business rules are chosen with an *intended* interpretation. However, the model may not incorporate an adequate nonobjectivist account of what makes the data model meaningful to the person(s) whose thinking is being modelled. According to Lakoff an adequate theory must take into account how the *content* of a concept is related to bodily experience.

3. THE EXPERIENTIALIST STRATEGY

The theory of categorisation based on cognitive psychology is an alternate approach for establishing a theoretical foundation to conceptual data modelling. In particular Lakoff (1987) describes an experientialist strategy, which has a central focus on meaning. The claim is that conceptual categories are very different from what the objectivist view requires of them. The experientialist approach to meaning contrasts to the objectivist approach whereby meaning is defined in terms of "our collective biological capacities and our physical and social experiences as humans functioning in our environment" (*ibid)*. The purpose of the strategy is to address the central problem with objectivist semantics whereby linguistic structures and the concepts they relate to are expressed as symbolic structures, meaningless until they derive meaning through interaction with things or categories in the 'actual' world (Lakoff, 1987). The strategy recognises that a large portion of our categories are not categories of 'things'; they are categories of abstract entities such as events, actions, emotions, governments, illnesses and social relationships.

The experientialist strategy understands experience in the broadest sense and characterises meaning by the "nature and experience of the species and of communities" (Lakoff, 1987). Experience in this sense is defined as everything that plays a role in the totality of human experience, for example genetically inherited capabilities, bodily nature, physical functioning and social organisation. The strategy is concerned with explaining why the human conceptual system (organisation of categories) is as it is. The claim is that features of the conceptual system arise due to the nature of our physical experience, derived originally through our interaction in a physical and social environment. Lakoff proposes two kinds of structure: basic level structure (refer section 3.3) and kinesthetic image schematic structure (refer section 3.4), these are products of our experience plus our ability to construct concepts especially those that fit with our innate experience.

3.1 Idealised Cognitive Models

If the activity of data modelling is 'knowledge creating' then it involves an organisation of knowledge by means of structures which Lakoff calls idealised cognitive models, or ICMs. Each ICM is described as a gestalt, comprising four kinds of structuring principles: a propositional structure, an image-schematic structure, metaphoric mappings and metonymic mappings. He describes a simple ICM using the English word *Tuesday*. He suggests that *Tuesday* can be defined according to an idealised model that includes "the week is a whole with seven parts organized in a linear sequence; each part is called a *day*, and the third is *Tuesday*" (*ibid*). Another example provided is the concept of the weekend, which he says, "requires a notion of a *work week* of five days followed by a break of two days, superimposed on the seven-day calendar" (*ibid*). He says that this model of the week is idealised, as seven-day weeks do not exist objectively in nature, but are created by people. But again, according to Bunge's ontology we cannot include such notions on a conceptual data model.

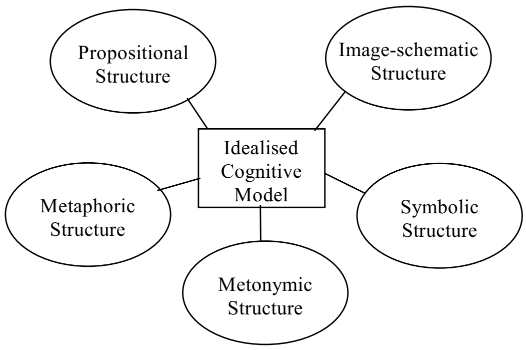

Figure 1: Structuring principles of ICMs (based on Lakoff, 1987)

Figure 1 represents the five structuring principles of ICMs. The first structuring principle (or model) is the propositional model; this model specifies the elements, their properties, and the associations amongst them. The example provided by Lakoff is a propositional model characterising knowledge about fire, which would include the fact that fire is dangerous. Whereas the second type of structure, image-schematic, specifies, "schematic images, such as trajectories or long, thin shapes or containers" (*ibid*). Metaphoric models are "mappings from a propositional or image-schematic model in one domain to a corresponding structure in another domain" p.114). Metonymy is the capacity to let one thing stand for another for some purpose, so a metonymic model can be either: social stereotypes, typical examples, ideal cases, paragons, generators, submodels or salient examples, that, according to Lakoff (1987) all produce prototype effects of some kind. A prototype effect is where certain members of a category are judged more representative of the category than other members (Lakoff, 1987). For example, "robins are judged to be more representative of the category BIRD than are chickens, penguins and ostriches" (*ibid.* p.41) .Like Putman's (1975) description of a stereotype as an idealised mental representation of a normal case, Lakoff defines social stereotypes as something that can be used to stand for a category as a whole, but sometimes these are recognised as not being accurate. An example of a social stereotype provided by Lakoff (1987) was Robins and Sparrows are typical birds (p.86). He remarks that these social stereotypes are "categories that function as stereotypes for other categories. An understanding of such categories requires an understanding of their role as stereotypes" (*ibid*).

Conceptual ICMs are characterised by Lakoff "independently of words and morphemes of particular languages" (p.289), however when linguistic elements are associated with conceptual elements in ICMs, a symbolic model results. He concludes this discussion by writing,

"linguistic expressions get their meanings via (a) being associated directly with ICMs and (b) having the elements of the ICMs either be directly understood in terms of preconceptual structures in experience, or indirectly understood in terms of directly understood concepts plus structural relations" (*ibid.*p.291).

If data modelling reflects some aspect of human reasoning, a corresponding cognitive semantic framework can be constructed using image schemas, metaphors and metonymies. Producing 'meaning' from a data model may be achieved by defining the components of the model according to Lakoff's schematic structures. Furthermore, these structures may be used to formalise the identification of semantic concepts described by Date (2000). Date (2000) claimed that semantic modelling includes, the identification of semantic concepts, the set of formal or symbolic objects, general integrity rules and finally the formal operators for manipulating the formal objects.

3.2. Concepts and Categories

In an information system the question of what a 'thing' is, is its category (Kent, 1978). We have categories of everything we think about. Categories are categories of things (Lakoff, 1987). We specify the set of categories to be maintained in an information system, these categories can differ from system to system. The classification rules used in one system to determine a category may be quite different in another system. The experientialist strategy maintains that the category of a thing may be determined by its position, environment, or use, rather than the set of defining attributes. Furthermore, bodily experience and how we use imaginative mechanisms determines how we create categories to understand experience.

Like Kent, Lakoff claims that there is no natural set of categories. Categories can be characterised using cognitive models of five types: propositional, image-schematic, metaphoric, metonymic and symbolic. Lakoff argues that cognitive semantics and the study of the general forms of metaphoric, metonymic and image-schematic reason can be applied to any subject matter that can be understood using image schemas, metaphors, and metonymies. He demonstrates cognitive semantics that covers the subject matter of predicate calculus, but then argues that the resulting logic would apply to any subject matter that can be understood in terms of these schemas and would provide an intuitively meaningful semantics. This requires that the logic would have an experientialist rather than an objectivist interpretation.

Lakoff describes the notion of concepts, for example the concept flight attendant can be characterised relative to an airline scenario (because for every concept there can be a corresponding category). A category is "those entities in a given domain of discourse that the concept (as described by the cognitive model) fits" (p.286). Concepts characterised in a cognitive model using necessary and sufficient conditions generate classically defined categories, whereas prototype effects can arise in a number of ways such as: metonymy, radial categories, generative categories and graded categories.

Cognitive models structure thought and are used when creating categories and in reasoning. According to Lakoff the structuring of concepts allows us to comprehend, to acquire knowledge and to communicate. The structuring of cognitive models however, does not account for meaningfulness. The experientialist strategy claims that conceptual structures are meaningful because they are embodied (tied to bodily experience). There are two kinds of structure in our preconceptual experiences: basic level and kinesthetic image-schematic structures: These concepts are directly understood in terms of physical experience

3.3. Basic-level Structures

This is the level of physical experience where we can distinguish tigers from elephants (human interaction with the external environment - the basic level), described through basic gestalt perceptions (overall shape), motor movements and the formation of mental pictures. Basic level concepts can be objects, actions or properties.

3.4. Kinesthetic Image Schematic Structures

Kinesthetic image schemas are a form of conceptual structure in cognitive semantics (Saeed, 1997). Lakoff maintains that image schemas structure concepts, that is, they define most of what we mean by 'structure' when we examine abstract domains. These image schemas structure our experience of space. The notion is that we form basic conceptual structures because of our physical experience of existing in the world (perceiving, moving, exerting and experiencing), this is how we organise thought. For example, the CONTAINER schema (or class) consists of a boundary which is the delineation of interior from exterior (the difference between in and out). Lakoff describes how we understand an enormous amount of activities in 'container' terms. For example personal relationships can be perceived in terms of containers: 'trapped in a marriage' and 'get out of'. Whereas the PART WHOLE schema represents how hierarchical structure is understood and can be metaphorically mapped onto subclass relations. IDENTIFICATION schema implies that all entities have a special property that identifies that entity. The LINK schema relates two entities that is the relational structure between them. However, the SOURCE-PATH-GOAL schema recognises that complex events have initial states (source), a sequence of intermediate stages (path), and a final state (destination). The path schema recognises that everyday experience involves moving around the world and experiencing the movements of other entities. This schema is perhaps most useful for describing events which are not usually represented in data modelling.

Table 1 provides an initial description of how the semantics of a data model may involve metaphorical mappings based on relevant image schemas. An understanding of how such metaphorical mappings are conceptualised by humans when creating data models may be useful for revealing different points of focus (alternative models) and for improving understanding and meaningfulness.

Schema	Metaphorical Mapping	Structural Elements	Description
CONTAINER	Class (the world is made up of entities). The underlying schema of everything.	INTERIOR, BOUNDARY, EXTERIOR	Each entity is represented structurally by a CONTAINER schema.
PART-WHOLE	Sub class. Type versus instance.	A WHOLE, PARTS and a CONFIGURATION	All entities will have certain properties in common.
IDENTIFICATION	Unique identifier.	IDENTITY	Every entity has a special property that serves to identify that entity.
LINK	An entity can be related to other entity by means of relationships.	Two entities, A and B, and LINK connecting them.	To secure the location of two things relative to one another. Represents how the relational structure is understood.
SOURCE-PATH-GOAL	Intension. (Purposes is understood by destination. Purpose is understood as passing along a path from a starting point to an endpoint).	A SOURCE, a DESTINATION, a PATH and a DIRECTION.	Complex events have initial states (source), a sequence of intermediate stages (path), and a final state (destination).

Table 1: Image Schemas for a Data Model

Lakoff says that each schema has an internal structure and may be understood in terms of direct experience. However, other more interesting forms of metonymic reasoning relevant to data modelling may be: social stereotypes, typical case, ideals, paragons, generators and salient examples. He remarks that these are normal activities involving the use of human reason, which involve “imaginative projections based on understanding an entire category in terms of some subpart of that category”. Also the conceptual categories “have properties that are a result of imaginative processes (metaphor, metonymy, mental imagery) that do not mirror nature” (p.371).

Using metonymy or reference-point reasoning a data modeller may determine intention. A metonymic model has the following characteristics:

- There is a “target” concept A to be understood for some purpose in some context.
- There is a conceptual structure containing both A and another concept B.
- B is either part of A or closely associated with it in that conceptual structure. A choice of B will uniquely determine A, within that conceptual structure.
- Compared to A, B is either easier to understand, easier to remember, easier to recognise, or more immediately useful for the given purpose in the given context.
- A metonymic model is a model of how A and B are related in a conceptual structure; the relationship is specified by a function from B to A. (Lakoff, 1987, pp.84-85).

These characteristics infer that B may be used to stand, metonymically, for A. Types of metonymy are; social stereotypes, typical cases, ideals, paragons, generators, and salient models. These may be useful classification techniques for supplementing the requirements definition phase of systems analysis. The data modeller may use metonymic models to aid understanding of the problem domain.

A social stereotype is when a judgement is made about people or a situation. Social stereotypes (according to the stakeholders perception) may be used to describe parts of the data model. However, Lakoff warns, “they are usually recognised as not being accurate, and their use in reasoning may be overtly challenged” (p.85). Therefore, recognising when a person is using a social stereotype to describe a category as a whole may help avoid placing too much importance on such descriptions. A typical category is when inferences from typical to atypical cases are made, based on knowledge of the typical. It is very normal to categorise things in terms of typical cases, however, a ‘good’ data modeller would try to model typical cases as well as those cases that were atypical.

A category can be understood in terms of abstract ideal cases which involve making judgements of quality and planning for the future. According to Lakoff a lot of cultural knowledge is organised in terms of ideals, for example ideal jobs, ideal workers and ideal bosses. Paragons include making comparisons and then using that comparison as a model for behaviour. Generators define concepts by principles of extension, where the members of a category are defined or generated by the main members. Salient examples include using familiar, memorable examples to understand categories. The data modeller may use many types of examples not only salient examples to understand the categories.

4 EXAMINING CATEGORISATION: A CASE IN QUESTION

A case study was undertaken to investigate the design and use of a complex data warehouse at a government department (Sampson and Atkins, 2002). In 1996/1997 a new system, AMS, was developed to record and manage the processing of all visa applications. The system was custom built using Microsoft SQL Server 6.5 relational database management system. At the same time the data warehouse was built to provide better access to management information, the rationale was to improve the accuracy of data for business monitoring. The source database (AMS) is very complicated due to the complexity involved with processing applications whilst adhering to legislative requirements. The information captured in the AMS system is the source of all information for the data warehouse (MIS).

The data warehouse was developed to meet the needs of three types of user groups: senior managers, analysts in the national office and branch managers.

The participants in the case study were: the data warehouse designer/developer, a senior consultant, the IT director, a technical support user and four data consumers. A number of interviews were carried out to identify factors which inhibit or prevent an in-depth understanding of semantic content of the data warehouse these were classified as: technical, social/cultural, training, resource, data or design related. In this paper I will focus on the problems with the design of the data warehouse, that is design decisions regarding entity and property decisions.

The most significant problem stemming from the underlying database design (AMS) is the fact that every document is handled as an application even where the documents quite obviously record different information. Participant G spoke of this problem from a design point of view,

"The data is an issue too, which is only now beginning to be sorted out, in that, it's the way they handle everything as an application. But different types of applications are very different like a visitors visa application goes through a very different process and requires different information to a permit issued over the border, or a ministerial appeal".

Furthermore, participant E says that a lot of the data on AMS is not carried onto MIS,

"There are occasions where you want to know the demographic characteristics of people approved under general skills category? You can't actually get that information on MIS …When you're looking at the individual record of data it's quite often AMS that you have to go to after you identify through MIS what it is you want".

However, he also concedes that the MIS was not originally designed with research requirements in mind, therefore a lot of the information he would like is not captured. Nevertheless, participant E commented that MIS is not comprehensive as it "would need to be people based …as opposed to applications based therefore, having everything about individuals easily accessible".

In addition, a major complexity in the applications system is the notion of grounds codes, which at the logical level appear simple, but at the physical level have been designed in such a way that it makes it difficult for the users to select the appropriate data. The use of structured codes in database design often causes flexibility problems and is generally not recommended. For example, in AMS (and MIS) a grounds_code comprises three parts: application category, application type and application criteria. The application category code identifies whether it is a visa, a permit, an appeal or a border application. The next level identifies the application type code (a type might be visitor, work, residence), below that at the lowest level is the application criteria code (general skills, family, humanitarian, family marriage). This structure requires (of the users) an understanding of these codes, so the correct records are selected when performing queries on the data warehouse. Participants G and F both commented that there were problems with the users understanding the complexity of the grounds code structure.

The problems experienced with dealing with complexity in this model may have arisen because of the particular view the designers had of the problem domain. The data model for the source database (and the subsequent data warehouse) was created with a focus on application as the primary entity-type, a view also reflected in the overall business strategy to improve application processing time and entry numbers into the country. However, as the users pointed out different types of applications are very different, requiring different sets of information to be collected and analysed. Problems that crept into the system included the ability to enter an application and an applicant twice as there was no way of matching multiple applications to the one person.

An alternative approach would be to relate the problem domain to the image schemas presented at Table 1. This may be achieved because we can understand the problem using basic level and image-schematic concepts because they are directly understood in terms of physical experience.

Schema (Lakoff (1987))	Metaphorical Mapping	Structural Elements	Mathematical idea
CONTAINER	In this domain, the world is made up of people out of the country and people in the country.	INTERIOR = individuals residing 'in' the country BOUNDARY = country border EXTERIOR individuals 'out' of the country	Set e.g. Person.
PART-WHOLE	Different individuals with particular identities requiring different types of legal documents	A WHOLE, PARTS and a CONFIGURATION	Factor (or decomposition) Person type Identity
IDENTIFICATION	A unique identifier to identify an individual (but one individual may have more than one identity in this domain).	IDENTITY (passport identification)	Person identity
LINK	The link between the person and her/his identity (or identities).	Person A and identity A, related using a surrogate key.	Correspondence, Chain Person identity 1 Person identity 2 In a 'perfect' world this should be a one to one relationship and hence redundant.
SOURCE-PATH-GOAL	The status of potential immigrant to application under consideration to processed applicant.	A SOURCE = person, a DESTINATION, = country A, PATH = procedure followed DIRECTION = application status.	Continuity Specific type of process to be followed depending on the type of application and specific set of data required for processing.

Table 2: Person/Identity focused schema

The schemas at Table 2 categorise the individual (or applicant) as the main entity-type through focusing on the container schema. The advantage with using this holistic approach is that it reveals the complexity between the individual and an individuals identity. This perspective recognises that there is an artificial difference between person and identity. A typical implementation solution would be to create surrogate keys for the individual (Simsion, 2001). Nonetheless this does not solve the real world problem of matching person A with identity A and person A with identity B. Moreover, it does not answer the philosophical questions:

- what is an identity?
- what distinguishes an individual from her/his identity?
- do two separate identities mean we have two separate individuals?
- is an identity specific to one individual?
- when does identity one become identity two? How do we capture the transformation? Is this necessary?

These questions are difficult to answer using traditional data modelling techniques. However, in this case they need to formally define the term identity. Furthermore it might be impossible to recognise a change in identity because in fraudulent circumstances the change may have a criminal element.

CONCLUSION

Lakoff (1987) says that image schemas and metaphorical models are required to represent the meanings of expressions. The meanings of the concepts in a data model can also be represented using image schemas and metaphorical models, this is because image schemas define what we mean by 'structure'when we talk about an abstract domain. The image schemas described above can be formally defined using logic but Lakoff argues that this is not necessary because the experientialist strategy claims that conceptual structures are meaningful because they are embodied (tied to bodily experience).

This research in progress claims that categories in general are understood in terms of Container schemas. However this research needs to be empirically tested to determine the efficacy of the strategy for data modelling. To test the use or otherwise of the experientialist strategy we suggest a number of experiments that firstly, test the use of the container schema for determining the difference between things and properties. Secondly an experiment which determines whether the notion of source-path-goal is relevant for distinguishing between entities and events. Thirdly test the understanding a group of users has of a data model created using the experientialist strategy and fourthly to examine the cultural implications of categorisation.

REFERENCES

Date, C.J. (2000): *An Introduction to Database Systems (7th Edition),* Addison-Wesley, Reading, Massachusetts.

Kent, W. (1978): *Data and Reality,* North-Holland, Amsterdam.

Lakoff, G. (1987): *Women, Fire and Dangerous Things*, University of Chicago Press.

Putman, H. (1975): The Meaning of "Meaning", in: GUNDERSON, K. (ed.), *Language, Mind and Knowledge*, Burns & Maceachern Ltd, Canada.

Saeed, J.I. (1997): *Semantics*, Blackwell Publishers, Oxford.

Sampson, J. and Atkins C. (2002): *Semantic Integrity in Data Warehousing: A framework for understanding"*, 35th Annual Hawaii International Conference on System Sciences (HICSS-35) January 7-10.

Simsion, G., (2001): *Data Modelling Essentials Analysis, Design and Innovation*, 2nd Edition, Coriolis, Phoenix, Arizona.

Wand, Y., Storey, V.C., and Weber, R. (1999): An Ontological Analysis of the Relationship Construct in Conceptual Modeling. *ACM Transactions on Database Systems, Vol. 24, No. 4*. 495-528.

Wand, Y., and Weber, R. (1993): On the Ontological Expressiveness of Information Systems Analysis and Design Grammars. *Journal of Information Systems*. 217-237.

Weber, R. (1997): *Ontological Foundations of Information Systems*. Blackburn Australia, Buscombe Vicprint.

ACKNOWLEDGEMENTS

I would like to thank Dr Robert Johnston, Senior Lecturer of the Department of Information Systems, University of Melbourne for his comments on this research. I would also like to thank Dr Clare Atkins, Academic Leader of the Nelson Marlborough Institute of Technology New Zealand, for her advice and support.