Constructing a Data Schema from an Information Flow Model

JUNKANG FENG
School of Computing
University of the West of Scotland
High Street Paisley, PA1 2BE UK
junkang.feng@uws.ac.uk

SUFEN WANG
Glorious Sun School of Business & Management
Donghua University
1882 West Yan'An Road, Shanghai 200051
China
sf_wang@dhu.edu.cn

Abstract: - The 'information content' of a data schema is concerned with the capacity of a database in representing the information that the database is designed to provide. It is recognized to be 'difficult to define and measure'. Our literature survey seems to show that this is an unsolved problem and the difficulties seem to lie with the lack of separation of information and data, and particularly with intuitive treatment of information. We examine what is required for solving this problem. We propose an approach to information and information flow for conceptual data modeling by drawing on a set of contemporary theories concerning the semantic aspect of information. With this approach, we formulate an information flow model from human purposeful activities from which to construct a data schema. This way it can be sure that the data schema represents required information, and therefore the latter is definitely in the 'information content' of the former. We observe that this constitutes a possible solution to this problem, and it also represents a 'semantic information theoretic' approach to conceptual data modeling. This work is a result of a substantial study of this problem including several real world case studies.

Key Words: - Information content, Conceptual modeling, Database design, Requirements engineering, Human purposeful activity

1 INTRODUCTION

A formal information system uses structured data to represent information. The structure of the data is specified with a data schema in one data model or another. A prominent problem with data schema construction is the 'information content' of such a schema, and it is recognised to be 'difficult to define and measure' (Batini et al., 1992, p.144). An extensive survey of the database design literature over 100 articles that we have conducted does not seem to show that this problem has been solved. The difficulties seem to lie with the lack of separation of information and data, and with intuitive treatment of information. Information and data are often treated as synonyms, for example, Date 1994, p4. A data schema is called an 'information model' (Shlaer and Mellor 1992, Flynn and Diaz 1996, Halpin 1995, p.5) or an 'information-structure perspective (ISP)' (Kahn 1985). Information is also defined as 'facts' (Mortimer 1993, p.7; Halpin 1995, p.5) and it is said that information must be informative and have meaning, that is, it informs (Mortimer 1993, p.8). But 'meaning' is perhaps more difficult to define (Stamper 1997).

In this paper some result of a substantial study (Feng 1999) of this problem is reported, which is a mechanism for the establishment and representation of the information content of a data schema, in conjunction with some of our more recent works (Feng 2002, Wang and Feng 2007, Wang et al., 2007). This work aims to make contributions to conceptual database design by making sure that a database is capable of representing the information that the database is designed to provide. This mechanism consists of the following elements:

• The idea of 'required information' as the thread linking the different stages in a process of constructing a (conceptual) data schema;

• A mechanism for formulating information flows and required information from human purposeful activities;
• A technique for systematically identifying data entities and other constructs for a data schema through analysing the usage of raw data;
• A technique for analysing the information bearing capacity of a database structure and database constraints specified by a data schema.

We concentrate on the fist two bullet points above in this paper as they are concerned with our basic approach. In section 2, we mention a number of related works. In section 3, we outline our approach and its foundations. In section 4, we describe the approach in some detail, starting with a set of basic concepts. Finally we conclude this paper with Section 5.

2 RELATED WORKS

We consider our research question in the context of conceptual modeling. Wand and Weber (2002) identify four elements of conceptual modeling: grammars (constructors), procedures (methods), scripts (models) and contexts. Maes and Poels (2007) consider the quality of conceptual modelling scripts from the user perspective. Ågerlalk and Eriksson (2004) identify that the 'static view' that conceptual modeling is concerned with emphasizes static properties in terms of entities and relationships. Weber (2003), and Seta et al., (2006) propose that conceptual modeling is an activity to build an idealized and simplified representation of selected semantics about some real-world domain. Ontology provides a perspective for considering grammars of conceptual modeling (Wand and Weber 1988, 2004; Weber 2003; Seta et al., 2006). Ågerlalk and Eriksson (2004) employ speech act theory (not ontology) as a foundation for conceptual modeling.

We observe that grammars (i.e., constructors) and the identification of the 'conceptual elements' (Andrade et al., 2006) for which constructors have to be provided are essential for conceptual modelling. For example, Bodart et al., 2001 discusses whether and when 'optional attributes and relationships' should be used as constructors in conceptual schema diagrams. We notice the 'orthogonal database design principle' (Eessaar 2006).

Drawing on these works, we define the 'information content' of a data schema to be the information that can be derived from the stored data defined by the schema. Information derivable from stored data is unlimited due to 'information nesting' (Dretske 1981, p.71). From the point of view of database design however, we want to make sure that the information that can be derived from stored data includes the information that the database is designed to provide. We will term the latter 'required information'.

However , little has been found in the literature on this problem of the 'information content' of a data schema. Eick and Lockemann (1985, p.88) define the concept of information preserving transformation between 'S-diagrams' (a variant of the binary relation model) as whether a correct S-diagram can be transformed into another correct S-diagram. Batini et al., (1992) suggest that a means for comparing the information content of two schemas is to compare their ability to reply to queries. Moody (1998) presents a set of metrics for evaluating the quality of entity relationship models, which does not explicitly cover the information content of such a model. The most relevant seems the 'completeness'. He defines four types of completeness errors, which seem neither formal nor quantitative. None of these works seems an adequate solution to the problem in question.

3 OUTLINE AND FOUNDATIONS FOR THE MECHANISM

To solve the 'information content' problem would require a mechanism, for which we also tap on wisdom in the literature. Pellens et al., (2007) discuss incorporating domain knowledge and high-level modelling concepts for describing virtual environment (VE). Schewe et al., (2005) look at user profiling and storyboarding in conceptual modelling of web information systems. Rolland and Prakash (2000) observe that the exploration of objectives of stakeholders and the activities that they carry out to meet these objectives is important for requirements engineering (RE) and 'RE product models use concepts for modelling these instead of concepts like data, process, event., etc., …'. Robinson (2008) looks at how to identify data requirements from a conceptual model.

We envisage therefore that our mechanism should fulfil three tasks. The first one is the identification and formulation of required information from human purposeful activities. The second is the derivation of a data schema from the formulated required information. And finally the third one is the analysis of a data schema in terms of whether the required information can be derived from stored data defined by the data schema. In the light of the afore-referenced works, this mechanism

is therefore concerned with 'grammars (constructors)' and 'procedures (methods)' of conceptual data modelling, but it is from an innovative perspective of representing semantic information with data.

To develop such a mechanism, the essential task is to separate information and data, to properly define them and the relationship between them. To this end, a set of theories concerning the semantic aspect of information is taken as the basis, which are Drestske's (1981) semantic theory of information, Mingers' (1995) framework of sign, information and meaning, Barwise and Perry's (1983) situation theory, Devlin's (1991) information flow theory and Floridi's (2005) information philosophy..Dretske's theory gives us the notions of the origin, quantity, and content of information, the concepts of 'informational relationship', and 'information content of a signal'. Mingers' (1995) framework clearly defines the interrelations between sign (including data), information and meaning. Mingers maintains that a sign carries information about states of affairs in the world – what it signifies, even though the sign may never be actually observed by anyone. Mingers defines three levels of meaning. Barwise and Perry advise us to look at how an agent divides the world up by using the ideas of 'real situation' and 'abstract situation'. Devlin provides a mechanism for modelling information flow, which makes use of a series of concepts, including 'infon', 'situation types' and 'constraints'. Floridi examines the alethic nature of *declarative*, *objective* and *semantic* (DOS) information, and argues that alethic neutrality that information is assumed to have by many is incorrect, and meaningful and well-formed data constitute DOS information only if they also qualify as 'contingently truthful'.

4 THE PROPOSED MECHANISM

In this section we describe the proposed mechanism in some detail. The mechanism consists of two parts. One is a set of basic concepts based upon the theories mentioned in the previous section. The other is concerned with capturing information flow within a human purposeful activity, and then identification of data representation of required information. In the description below, material extracted from a substantial case study on a property leasing company – the Cleland and Fleming Company (C&F for short) – will be used when and as required.

4.1 Basic concepts

4.1.1 Origin, quantity and content of information

Following Dretske (1981), information will be taken as created by or associated with a state of affairs among a set of possible outcomes of a selection process, the occurrence or realization of which reduces uncertainty. The quality of the information that is created by or associated with a state of affairs is the actual reduction in uncertainty, which can be measured by using probability theory. Information is therefore seen as an objective commodity. The content of information may be seen as a state of affairs (a situation) and what is or will be true in that state of affairs.

4.1.2 Information is carried by signals

A state of affairs, say r_1, is one of possible outcomes of a selection process, say r. The reduction in uncertainty at r due to the occurrence of r_1 may be accounted for by one or more events, say s_1, s_2, ..., s_n, that occur at another selection process, say s. This gives rise to a special kind of relationship - 'informational relationship' (Dretske 1981, p.35) between r and s. An informational relationship captures certain degree of dependency between a state of affairs r_1 of r and what takes place at s. This dependency can be demonstrated by the fact that r_1's appearance alters the distribution of probabilities of the various possibilities at s. The dependency is a type of regularities concerning different selection processes that are based upon nomic dependencies (Dretske 1981), logic, or norms, etc. in a social setting.

Due to this relationship, information created at s carried by r. Thus s is the 'information source', and r 'the carrier or bearer of information' about s. For example, a state of affairs r_1 at r can be seen as a signal that carries information about s in terms of what state s is in. Moreover, if it is recorded, r_1 becomes a piece of data. Thus data carry information. In general, data in a database are a collection of recorded signals or events that carry and therefore provide information about a real world domain.

4.1.3 An agent's acquiring, recording and sending information

To establish the information content of a data schema, a basic task is to find out how an agent acquires, records and sends information. In their purposeful activities and in general their maintaining relationships with others in a social

setting, an agent continuously learns about situations relevant to him/her and communicates with others. That is, the agent acquires (which can be divided into 'receiving' and 'obtaining') information and forms intentions to take actions. Information acquisition can take place only because there is some informational relationship between a signal and an information source. The information about the source that a signal carries and all other information nested in it are what we call the 'informational content' of the signal. A signal may carry more information than an agent can actually receive. And different agents may receive different amount of information and indeed different information from the same signal. To capture this, we borrow Dretske's idea and define the notion of relative information content of a signal as 'an agent gets the information that *s* is F from signal *r* if the conditional probability of *s*'s being F, given *r* (and k), is 1 (but given k alone, less than 1), where k is how much the agent knows about the possibilities at the source'. Thus k captures this 'relativisation' of information contents of a signal for different agents.

Mingers' 'meaning system' (1995) for an agent can now be seen as made up of three elements. The first is the collection of a variety of relative informational contents of various signals. The second is all other information that is nested in the 'direct' relative informational contents. Information nesting is defined by Dretske (1981, p.71) as 'the information that *t* is *G* is nested in *s*'s being *F* = *s*'s being *F* carries the information that *t* is *G*'. The third is his/her intention to take actions after having received the first, and obtained the second.

An agent also records information, which creates data. Recording information can be achieved by doing one of the following. First of all, one can record the signal or observed event, which will bear all the three levels of meaning. To retrieve any of them, a process of interpreting the recorded signal or event appropriate to the level required will have to take place. Secondly, one can record a specific piece of meaning in one of the three levels. The higher the level and the more specific the piece of meaning is, the smaller the scope of the acquired information is that will be borne by the data. When we design the data structure for an information system, this issue should be taken into consideration.

Furthermore, an agent sends information to others. Following Mingers' (1995) idea, sending off some information can be looked at by using a reverse process of an agent's acquiring information. That is, from the third level - an intention to take an action, to the second level - nested information, to the first level - information directly carried by a signal (and not implied by any other information), and then to the signal that carries the first level meaning. Depending on how much the sender believes the receiver knows how to interpret it, a signal will be chosen to carry information in one of these levels. Again, the higher the level and the more specific the piece of meaning is, the smaller the scope of the acquired information is that will be borne by the signal to be sent.

Now we introduce a set of constructors for formalising information and the mechanism of an agent's acquiring, recording and sending information.

4.1.4 Items of information

We said earlier, the content of information may be taken as a state of affairs. A state of affairs can be seen as made up of one or more primitives, which can be expressed as a number of individuals having or not having certain relationship or property at a temporal location and a spatial location. So information is made up of items, each of which consists of two parts – a statement that some particular individuals posses or do not posses a certain property or relationship, and a context within which the statement is true. These all are intuitive terms. We now formalize them.

First, we will use a formal concept 'infon' (After Devlin 1991, p.22) to model the 'statement' by using a predicate:

$$r(a_1, \ldots, a_n, l, t, 1),$$

which means that individuals $a_1, \ldots, a_n$ have property or relationship *r*, at temporal location t and spatial location l. The last argument 1 in the above predicate expression is one of the two possible Boolean values that the polarity may have. For example,

makesenquiry(Jane Smith, 24/3/08, 1)

is an infon, which means that Jane Smith makes an enquiry about leasing a property on 24th March 2008.

The elements in an infon are called arguments. When all arguments are constants or bound variables, the infon is said a 'parameter free infon', otherwise a 'parametric infon'. The latter is a template for the former. For example,

makesenquiry(client', 24/3/08, 1)

is a parametric infon as client' in it is an unbound variable (we will always use a '' to indicate a

variable in this paper). An unbound variable can be assigned a constant (called 'anchoring' by Devlin (1991, p.134)) in a particular situation.

Second, an infon is only true in a certain context - a real situation. For example, the above infon is only true in the situation where a client makes an enquiry about properties for lease in May 2008. We will use the formal concept 'abstract situation' ('situation' for short) to model the term 'context'. An abstract situation is the context in which a set of infons is true. If the above infon denoted with, say, σ, is true in a situation s, then we write

$$s \models \sigma.$$

The relationship between a real situation and its corresponding abstract situation is

$$s_a = \{\sigma \mid s_r \models \sigma\}$$

where s_a is an abstract situation, s_r is a real situation, and σ is a set of parameter free infons. Moreover, the formal concept 'situation type' is a set of abstract situations, and any situation is an instance of a situation type. For example,

$$S_1 = [s_1' \mid s_1' \models \text{makesenquiry(client', C\&F, enquirydate', 1)}],$$

is a situation type, which is a collection of abstract situations in each of which a client makes an enquiry at the Cleland and Fleming Company. We suggest using the term 'info unit' to refer to the combination of a situation and the infon(s) that are true in the situation.

4.1.5 Information flow

We use 'information flow' to formalize the intuitive term of 'an agent receives information from a signal or event,' and 'an agent obtains information from some other information.' The latter means that an agent obtains some information that is nested in the information that he/she already possesses. We will formalise an information source by using a situation type. The content of information received and obtained is state of affairs (we said this earlier), which is an instance of a situation type. Moreover a signal is also a state of affairs, so it can also be formalised to be a situation type. Therefore, the mechanism for information flow to take place can be seen as a directed connection between two situation types, which we call 'constraint' following Barwise and Perry (1983, p.119), Devlin (1991, p.12) and Barwise and Seligman (1997, p.29).

Now we will use the following scenario to illustrate the concept of 'information flow':

When a clerk at Cleland and Fleming Company sees that the name 'Jane Smith' is in the enquiry list and the enquiry date is 24th March 2008, the clerk knows that Jane Smith makes an enquiry about leasing a property on that date. That is, the former state of affairs carries information about the latter, and the clerk gets it.

We wish to formalise the above process. There are two situation types involved. For the information source, namely a client makes an enquiry, we can have

$$S_1 = [s_1' \mid s_1' \models \text{makesenquiry(client', C\&F, enquirydate', 1)}].$$

For the signal, namely a name appears in the enquiry list, we can have

$$S_2 = [s_2' \mid s_2' \models \text{inenquirylist(clientname', enquirydate', 1)}],$$

which is a collection of situations in each of which a client name and enquiry date appear in the enquiry list. We then define a constraint

$$S_2 \Rightarrow S_1,$$

which is a mechanism for the C&F clerk to obtain the information. This constraint exists because of how the job is done at C&F, which establishes an informational relationship between the two situation types, and the clerk in question can make use of it. This mechanism works like this: for S_2, if an individual situation is found as the clerk does in the above scenario where parameter clientname' anchors to the name 'Jane Smith', and parameter enquirydate' to 24th March 2008, which gives a certain state of affairs of the signal, then a certain affairs of the information source will be found where parameter client' in S_1 anchors to Jane Smith, and parameter enquirydate' to 24th March 2008.

A couple of points are in order. First, the state of affairs that Jane Smith makes an enquiry on 24th March 2008 creates information because there were many other possibilities of someone making an enquiry, and the uncertainty caused by these possibilities is reduced by this state of affairs. Second, the state of affairs that the name 'Jane Smith' appears in the enquiry list carries this information because given it the probability of Jane Smith's making an enquiry on that date is 1, otherwise less than 1.

In general, an information flow is a formulation of an agent's handling information. When all relevant information flows for a human activity are identified, we have an 'information flow model' for the activity. This model captures what and how information is received and obtained and from where. The information to be received and obtained is what we called 'required information' earlier, and it is the collection of the situation types in the second position of the constraints that gives us the types of required information. That is, in the above

example, S_1 is a piece of formulated required information. Note that this collection includes both situation types and infons. If it is a formalized signal or event and it is recorded in some way, then the situation type in the first position of a constraint becomes data that bear required information. In the above example, S_2 is a piece of formulated data that represents (carries) the required information S_1. So an information flow model also captures data and what an agents records and why with precision and formality. Moreover a piece of data can be sent as a signal, for example, S_2 might have been sent to the clerk by another clerk at C&F. Thus an information flow model also captures a signal that an agent sends and the information the signal carries with the same level of precision and formality.

4.2 Formulating required information in human purposeful activities

Using the basic concepts developed above, to formulate the information perspective of a human activity is a matter of identifying relevant situation types and constraints that connect them. It is found from our experiments with case studies (documented in Feng 1999) that the following situation types appear common:

The embedding situation – it refers to the immediate environment of the activity, which consists of the embedding situation before action and embedding situation after action. The former incorporates the 'driving force' for the activity and the information that is processed by the activity and the latter the direct result of the activity.

The basis situation - for an activity to be carried out, often some information is required as a basis or supporting material. This kind of information is not the immediate 'driving force' of an activity, not processed by it, not change through it and not a result of it. The basis situation can be seen as the embodiment of the 'k' in our definition of the 'relative information content' of a signal. The anchored infons in a basis situation form a condition for the agent to obtain information from the embedding situation before action.

The data storage situation - to anchor parameters within infons in an embedding situation and/or a basis situation, an agent may need to refer to some data. That is, data carries information about the embedding situation and the basis situation. We can describe this by defining a data storage situation and its linkage with the embedding situation and the basis situation. A data storage situation may change after some action having been taken.

The connections and relationships between these situation types are illustrated in Figure 1 below.

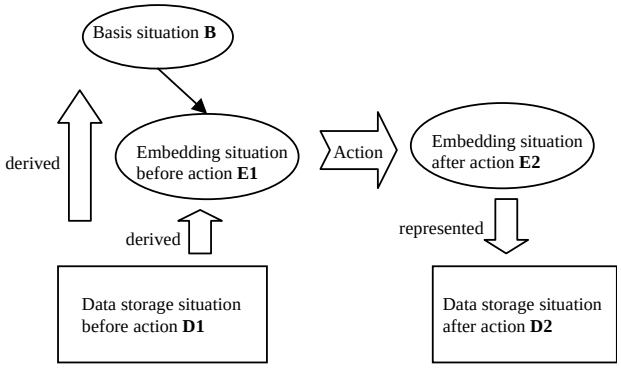

Figure 1 Relationships between basic situations in an activity

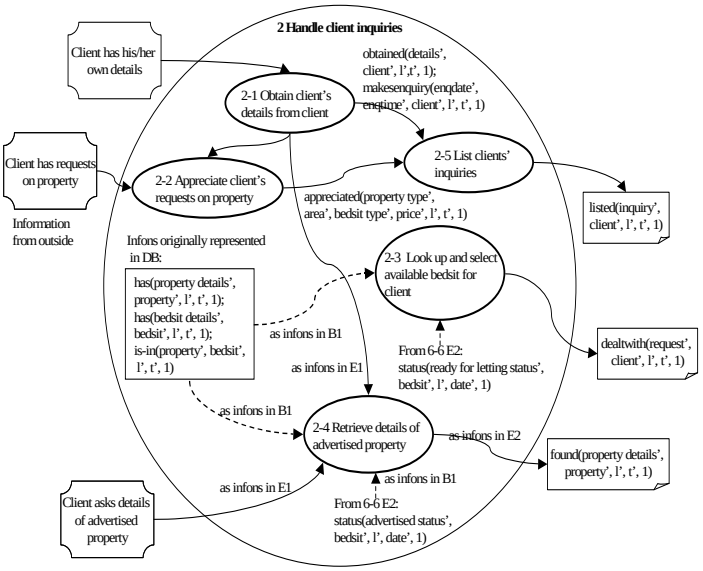

Figure 2 An information flow model for a complex activity

By using these situation types, an information flow model for a complex activity can be created, which captures the relationships between constituent activities from an information flow perspective. Figure 2 is an example taken from our case study on the C&F Company.

4.2.1 Constructing a data schema from raw data

We call the arguments in the infons of the data storage situation 'raw data' as they are raw material for constructing a data schema that should be

capable of representing required information. We need to structure these arguments into a data schema. The idea is to formulate and then analyze the usage of the raw data items in relation to elementary activities in order to classify them and find the relations between them. To this end, the raw data items need to be consolidated first with a view to making them mutually exclusive. Note that raw data are parameters, which are variables of certain types. And any type can be defined by using infons and situations as we did for situation types earlier. So the consolidating process is that of analyzing infons and situations. Then the data are classified in terms of how widely it is used and how independent it is. We use a set of usage parameters to quantify these. The usage parameters are task usage (TU), joint usage (JU), usage ratio (UR), and joint usage ratio (JUR). TU quantifies how widely a raw data item is used by activities, and JU how widely it is used with other raw data items. The independence is measured by the ratio between the number of other raw data item with which a raw data item is used in most tasks and the total number of other raw data item with which the raw data item is used, which is UR. JUR is defined to describe how close a raw data item is to another raw data item or a group of other raw data items. Through calculation of the parameters, relatively widely used and independent raw data items are classified as primary data, which lead to entities, and the rest are auxiliary data leading to attributes. Moreover, the attachment of an attribute to an entity is decided on JUR. As a result of these, data level constructs will emerge, and therefore a data schema will be formed. Details of this part of the work can be seen in Feng 1996. The data schema in entity-relationship model for C&F is shown in Figure 3.

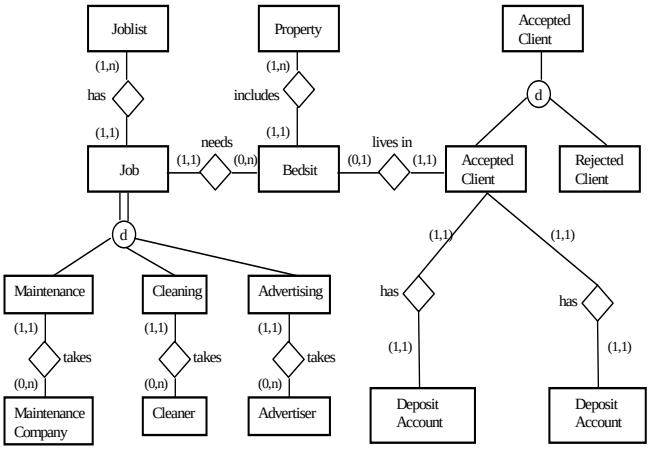

Figure 3 A data schema for the C&F Company

4.2.2 Analyzing the information bearing capacity of a data schema

The data schema formed through the previous stage is a preliminary and crude one in terms of whether it indeed represents the required information. The final stage of the proposed mechanism is to analyze and make sure that a quality schema be constructed. This requires looking at a data schema's 'information bearing capability' (IBC for short). We define a principle for IBC as follows:

A) For a token level data construct (or a 'media construct' in general), say p_j, or a collection of token level data constructs (or a collection of 'media constructs' in general), say $\{p_i,\ldots p_j\}$, to be capable of *representing* an individual real world object or an individual relationship between some real world objects (or a 'referent construct' in general), say i_i, which is neither necessarily true nor necessarily false [1], the following two conditions would seem sufficient and necessary.

1) Information Content Containment

The concept of 'information content' of a sign/message ('state of affairs' in general) can be defined as follows (following Dretske (1981, P.45)):

> *'A state of affairs contains information about X to just that extent to which a suitably placed observer could learn something about X by consulting it.'*

For a data schema (a formal information system in general) to be able to bear a type of information regarding some particular information source, there must exist at least a path or a collection of paths in the data schema (or system) such that the information content of the possible instances of the path or collection of paths include all possible instances of the type of the information. The simplest case of the above is that the literal or conventional meaning of the path or collection of paths includes the concept or concepts which can be instantiated by instances (i.e., paths on the data token level) of the paths, and as a result of which all the individual pieces of information of the type of

[1] See Floridi 2005.

information in question is covered (i.e., included) by the information content of the instances of the paths.

To express the above formally,

Let I be the type of information to be represented, and i_1, i_2, …, i_i,… be individual pieces of information[2] of type I;

Let P be a path, and p_1, p_2, …, p_j,… be instances of P, i.e., data token level constructs of P;

Let IC(p_j) be the information content of p_j;

Let LitM(p_j) be the literal or conventional meaning of p_j;

For P to be able to represent I, the concepts in P and the structure and constraints of P must be such that for every i_i there is at least one p_j or a collection instances $\{p_i,\ldots\ p_j\}$ such that $i_i \in$ IC(p_j) or $i_i \in$ IC($\{p_i,\ldots\ p_j\}$).

The simplest situation of the above is where $i_i \in$ LitM(p_j) or $i_i \in$ LitM($\{p_i,\ldots\ p_j\}$).

Note that it goes without saying that the conditions regarding 'information amount' [3] must be satisfied in order for this condition to hold.

2) Distinguishability

Let Y be the system that stores and manipulates P and its instances p_1, p_2, …, p_j,…, among others,

The structure and constraints of P are such that p_j or $\{p_i,\ldots\ p_j\}$ is distinguishable from the rest of possible instances of P by the only means available to Y.

B) For a token level data construct (or a 'media construct' in general), say p_j or a collection of token level data constructs (or a collection of 'media constructs' in general), say $\{p_i,\ldots\ p_j\}$, that are capable of *representing* an individual real world object or an individual relationship between some real world objects (or a 'referent construct' in general), say i_i, to be capable of *actually providing* information about i_i, the following two conditions would seem sufficient and necessary.

3) Accessibility

p_j or $\{p_i,\ldots\ p_j\}$ must be accessible by the only means available to system Y.

4) Derivability

In the case where p_j or $\{p_i,\ldots\ p_j\}$ has neither literal nor conventional meaning and in the case where neither the literal nor the conventional meaning of p_j or $\{p_i,\ldots\ p_j\}$ is i_i, the user must be provided with a means by Y to infer i_i from p_j or $\{p_i,\ldots\ p_j\}$.

Note that in relation to condition of Information Content Containment above, for a particular situation (i.e., an information source, say S), the existence or occurrence of p_j or $\{p_i,\ldots\ p_j\}$ results in an alternation of the probability distribution of the possibilities of S. Moreover, with p_j or $\{p_i,\ldots\ p_j\}$ the probability of i_i is 1, and without p_j or $\{p_i,\ldots\ p_j\}$, is not 1. In relation to Condition 2 above, p_j or $\{p_i,\ldots\ p_j\}$ must be a distinct and distinguishable state of affairs among more than one possible state of affairs.

We believe that this principle is scalable to suit more (i.e., p_j or $\{p_i,\ldots\ p_j\}$ could be an instance of an entire system among many related systems, for example) or less (e.g., an instance of a simple attribute of an entity in an ER schema) complex cases, and hence flexible in terms of applicability.

Following the principle, in order for a data schema to be able to bear formulated required information, it would be sufficient if:

Every object type, of which one or more parameter is, found in the formulated required information is represented by one or more entity in the data schema; and

Every instance of an info unit in the formulated required information is represented by at least one structure such as a path in the data schema.

To this end, we analyse the primary meaning (i.e., the semantic content (Mingers 1995)) and the implied meaning of the constructs of a data schema against the info units described earlier. We have developed the notion of 'classes of a path' in an ER schema whereby whether an info unit is included in the meanings of a path can be decided. This part of the work requires much space to present and is thus beyond the scope of the current paper. Interested readers are referred to Feng and Crowe 1999.

4.3 Evidence of the usefulness of the mechanism

Through two substantial case studies documented in Feng 1999, part of which are cited in this paper, it emerges that this mechanism enables a data schema

[2] A type of information can be formulated as a parametric infon (Devlin 1991)

[3] See Feng 2002 for details.

to be constructed through a systematic procedure. Every step of the procedure is well defined and justifiable. In the process, formulated 'required information' plays a central role, the starting point and end. This increases the certainty of the data schema's informational correctness, completeness and minimality. In addition, arbitrary decisions related to 'non-determinism' (Hawryszhiewycz 1991, p.119) on choosing a modeling construct is reduced, and 'connection traps' (Howe 1983, p.113) can be identified and avoided by using the notion of 'classes of a path' (Feng and Crowe, ibid.).

5 Conclusions

This paper is concerned with the problem of the 'information content' of a data schema for databases. It is a crucial task and does not seem to have been adequately addressed in the literature. The difficulties seem to rest with the lack of separation of information and data, and intuitive treatment of information. In the overarching context of conceptual modeling, we have explored the grammars (constructors) and procedures (methods) of conceptual data modeling with a semantic information theoretic perspective. We have presented an approach that is based properly upon contemporary theories regarding the semantic aspect of information and information flow. We reported a mechanism whereby an information flow model is formulated from human activities based upon which a data schema can be derived. This way, a data schema that has required information content is guaranteed. Experiments with case studies gave positive supporting evidence that this mechanism is able to improve database design. This helps achieve the completeness, correctness and minimality of a data schema. In the process of doing so, the mechanism also alleviates difficulties related to well-known problems of 'non-determinism' and 'connection traps' in database design.

The ideas presented here can be, we envisage, applied to looking at following problems. One, how we can link a conceptual model created by using the Soft Systems Methodology and information systems design; Two, whether the capacity of a database in using data to provide information is formalizable and if so, how; Three, how the underlying mechanism for an information system to be useful for its targeted user may be uncovered and formulated.

References:

[1] Ågerlalk, Pär J.; Eriksson, Owen.(2004) Action-oriented conceptual modeling. European Journal of Information Systems, Mar2004, Vol. 13 Issue 1, p80-92.

[2] Andrade, J.; Ares, J.; García, R.; Pazos, J.; Rodríguez, S.; Silva, A.(2006).Definition of a problem-sensitive conceptual modeling language: foundations and application to software engineering. Information & Software Technology, Jul 2006, Vol. 48 Issue 7, p517-531.

[3] Barwise J and Perry J (1983) Situations and attitudes, Bradford Books, MIT Press.

[4] Batini C, Ceri S and Navathe S B (1992) Conceptual database design: An entity-relationship approach, Benjamin/Cummings, Redwood City in Calif.

[5] Date C J (1995) Introduction to database systems, 6th ed., Addison-Wesley, Reading, Massachusetts.

[6] Devlin K (1991) Logic and information, Cambridge University Press, Cambridge.

[7] Dretske F I (1981) Knowledge and the Flow of Information, Oxford, Basil Blackwell.

[8] Eessaar, Erki. (2006). Guidelines about usage of the complex data types in a database. WSEAS Transactions on Information Science and Applications, v 3, n 4, April, 2006, p 712-719.

[9] Eick C F and Lockemann P C (1985) Acquisition of terminological knowledge using database design techniques, Proc ACM SIGMOD 1985 International Conf on Management of data, SIGMOD Record Vol 4, No4 p84.

[10] Feng J (1996) Can 'entity identification' be disciplined? Proceedings of 5th International Conference on Information Systems Development (ISD'96), Gdansk, Poland, 24-26 Sept, 163-174.

[11] Feng J (1999) An information oriented approach to the construction of a conceptual data schema, PhD Thesis, University of the West of Scotland, UK.

[12] Feng J and Crowe M (1999) The notion of 'classes of a path' in an ER schema, Proceedings of 3rd East-European Conference on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, 13-16 September.

[13] Feng J., 2002: 'The 'Information Quantity' Aspect of the Information Bearing Capability of a Conceptual Data Schema', In Proceedings of the 7th Annual Conference of the UK Academy for Information Systems, UKAIS'2002. Leeds. ISBN 1-898883-149. pp.150-157.

[14] Floridi L., 2005. 'Is Semantic Information Meaningful Data?' in *Philosophy and Phenomenological Research* Vol. LXX, No. 2, March 2005.
[15] Flynn D and Diaz F (1996) Information modeling: An international perspective, Prentice-Hall Inc Englewood Cliffs, New Jersey.
[16] Halpin T (1995) Conceptual schema & relational database design, Prentice-Hall Inc Australia.
[17] Hawryszhiewycz I T (1991) Database analysis and design 2nd ed., MacMillan publishing company.
[18] Howe D R (1983) Data analysis for data base design, Edward Arnold.
[19] Kahn B K (1985) Requirement specification techniques, in Yao S Y (eds.) Principles of database design, Vol 1 Logical Organizations, Prentice-Hall.
[20] Maes, Ann; Poels, Geert. (2007). Evaluating quality of conceptual modeling scripts based on user perceptions. Data & Knowledge Engineering, Dec2007, Vol. 63 Issue 3, p769-792
[21] Mingers J (1995) Information and meaning: foundations for an inter-subjective account, Information Systems Journal, 5, 285-306.
[22] Moody D (1998) Metrics for evaluating the quality of entity relationship models, Proceedings of 17th International Conference on Conceptual Modeling (ER'89), Singapore, 16-19 November.
[23] Mortimer A (1993) Information structure design for databases - A practical guide to data modeling, Butterworth-Heinemann ltd.
[24] Pellens, Bram; De Troyer, Olga; Kleinermann, Federic; Bille, Wesley.(2007) Conceptual modeling of behavior in a virtual environment. International Journal of Product Development, 2007, Vol. 4 Issue 6, p6-6.
[25] Seta, Kazuhisa; Koyama, Kazuya; Hayashi, Yusuke; Ikeda, Mitsuru. Building ontologies for conceptual model management. WSEAS Transactions on Information Science and Applications, v 3, n 3, March, 2006, p 546-553
[26] Schewe, Klaus-Dieter; Thalheim, Bernhard (2005). Conceptual modeling of web information systems. Data & Knowledge Engineering, Aug2005, Vol. 54 Issue 2, p147-188.
[27] Shlaer S and Mellor S J (1988) Object-oriented systems analysis: modeling the world in data, Yourdon Press, Prentice Hall Building, Englewood Cliffs, New Jersey.
[28] Stamper R (1997) Organizational semiotics, in Mingers J and Stowell F (eds.) Information systems: An emerging discipline? McGraw Hill, London.
[29] Wand, Y. and Weber, R (1998). "An Ontological Analysis of Some Fundamental Information Systems Concepts," in Proceedings of the 9th International Conference on Information Systems, Minneapolis, MN, Nov 30-Dec 3, 1988
[30] Wand Y, Weber R. (2002) Research Commentary: Information Systems and Conceptual Modeling - A Research Agenda. Information Systems Research, Vol.3, No.4, pp.363-376.
[31] Weber, Ron. (2003) Conceptual Modeling and Ontology: Possibilities and Pitfalls. Journal of Database Management, Jul-Sep2003, Vol. 14 Issue 3, p1-20.
[32] Wand, Y; Weber, R (2004). Reflection: Ontology in information systems. Journal of database management, Vo.15 Issue: 2, pp.III-VI, 2004.
[33] Wang S, Lin C, Feng J. (2207) A Hermeneutic Approach to the Notion of Information in IS. Proceedings of the 7th WSEAS international conference on simulation, modeling and optimization, 2007, pp.400-404.
[34] Wang, Y; Feng, J (2007). FCA assisted IF channel construction towards formulating conceptual data modeling. WSEAS Transactions on Systems, v 6, n 6, June, 2007, p 1159-1167.
[35] Weber, Ron. (2003) Conceptual Modeling and Ontology: Possibilities and Pitfalls. Journal of Database Management, Jul-Sep2003, Vol. 14 Issue 3, pp.1-20.