

A Semantic Approach to the Notion of *Representation* and Its Application to Information Systems

Lin Liu

Database and Semantic Web Research Group
School of Computing, University of the West of Scotland
lin.liu@uws.ac.uk

Junkang Feng

School of Computing, University of the West of Scotland
Junkang.feng@uws.ac.uk

Abstract-Representation is a key concept for information systems (IS). IS is designated to provide information. However, the most available information systems such as databases using a computer are actually data storage and management systems because all they handle are data, which may be used to represent information. Therefore, a good understanding on the relationships between data, information and *representation* is vital for our study of information systems. And yet, it would seem that in the literature of information systems, the notion of representation has not been well studied, and in particular, how a possible representational relationship between two sets of data constructs may be systematically identified and how this representational relation may be explored for bearing on IS problems would seem worth looking at. The results of our investigation show that representation is a special type of signs created or being used by human agents to carry information about other states of affairs. Any representational relationship is realized and underpinned by information flow between components within an information channel. Moreover, representation can be quantitatively measured as long as certain conditions are met. The information content of representation reveals what you could possibly learn from it and the semantic content of representation is the proposition that a human originator intended to convey by using the representation. In addition, schema connections (i.e., connections between elements within a data schema) can be identified by examining the representational relationship between data entities.

Index Terms—Representation; Semantic content; Information Content; Information Flow; Schema connection

1. Introduction

What is *representation*? We follow Shimojima [23], and preserving the double usage of the English term we will use “representation” not only to denote a relation between two objects, in which case, ‘representation’ is always of a singular form as an uncountable noun, but

also to denote the thing that represents, namely, the object that has the relation of representation to the other object in which case ‘representation’ is a countable noun and can be plural.

In different disciplines, the notion of representation has been approached differently [3] [11] [23]. We observe that a representational relationship between two things takes place only because one of them carries the information that the other exists. We further observe that such two things have to be such that they can be formulated as states of affairs, and furthermore for a representational relationship to exist it has to be the case where a state of affairs is being used by human agents to carry the information that another particular state of affairs exists. There are two direct implications from this observation. Firstly, the creation of representation generally involves *cognitive states*. Secondly, no representation exists without information flow. We believe that actually it is information flow within a distributed system [3] that enables representational relationships between its constituting components.

With our approach, a good understanding of information is necessary for defining the notion of representation. We adopt the philosophical stand of semantic information that information is an objective commodity, most of which is carried by signs/signals/states of affairs in analogue form [11, p139], but information is independent of its carrier and receiver. The creation of information is due to reduction in uncertainty of what might have happened. Despite the human’s involvement in the creation of certain types of representation, in this paper, we only consider physical representations and representation systems, and not mental or cognitive representations.

Representations can be seen as a kind of ‘information carrier’. Dretske defines *information carrying* as “what information a signal carries is what it is capable of ‘telling’ us truly about another state of affairs [11, p.64]”. How much information is generated and carried can be measured as long as the reduction in uncertainty, i.e., the probability of the event whose occurrence results in the reduction, is known. We propose that representation can be quantitatively measured in terms of how much

information is created and then transmitted from the represented (i.e., the information source) to a representation (i.e., the information carrier).

The study of semantics and semantic content is becoming an increasingly significant subject in IS applications because of its close association to user's interpretation of data. We argue that a database system is a kind of representation system in the way that data are being created and used by human agents to represent things in the real world. The semantic content of representation crucially reveals what the human originator's proposition or intention is by creating the representation. An accurate representation would carry such a proposition to its receivers (users) through the user's digitisation (interpretation) of the information that the representation carries. The semantic content of representation can be identified as what a representation is capable of 'telling' us truly about the representation *per se* plus the represented state of affairs.

Databases are one of key technologies employed for information systems. The development of databases could face enormous challenges due to the rapid social and economic changes. The fundamental challenge of database development is how to improve the efficacy and efficiency of a database so that it is able to provide information more precisely to its users. It would appear that the theory of representation is highly relevant to some theoretical as well as practical issues of databases such as database integration and reasoning, knowledge representation and management, semantic web services and connection traps among others.

Our journey on the notion of representation starts from philosophical analysis of some fundamental notions which is detailed in section 2. Then we move our attention to the main part, which is the theoretical approach on the notion of representation listed in section 3. We demonstrate how our representation framework can be utilized to solve IS problems, i.e., the connection traps in a database schema, which is outlined in section 4 and finally we give conclusions in section 5.

2. Some Fundamental Notions Related to Representation

It would be impossible a task to try and define the notion of representation for everyone to accept. The literature seems to show that the exploration of representation is carried out in two streams: the mode and the content. Research related to human mind falls in the former stream. Our interest is to develop an approach to representation concentrating on the content of a representation. As our approach is based on the notion of 'information', we have to start our endeavour with that of *information*.

2.1 The Notion of Information

As one of the oldest subjects of mankind, the notion of information has received many different definitions such as:

- Information is data that has been processed in some way to make it useful for decision makers, which is revealed by Lewis's [18] survey of 39 IS texts
- Information equals data plus meaning [6, p.303]
- A difference that makes a difference [4, p.286]
- Information is the propositional content of a sign [11, p.65], [19, p.6]

It seems unwise and impossible to gain a tacit agreement on information. It may be argued by Semiotics that information must be embraced as a material object, as an individual cognitive effect, and as a social institution. However, our ontological stand on information follows Dretske's objective view. Information is an objective commodity that exists independently of its creator as well as its observer, if any. The creation of information is the result of reduction in uncertainty.

Despite the connection between a sign and cognitive agents (e.g. human beings) in the social world, and despite the abilities of cognitive agents in generating information through signs, e.g., traffic signs, the making of the sign is independent of its observer if any. Also, information is not created in the mind of the observer of the sign. For example, if the utterance of a speaker is out, the information is there no matter who receives it.

Information can be acquired if the receiver is aware of and attuned to some constraints [8, p.15]. For example, a constraint concerning a tree trunk could be 'Number of rings' \Leftrightarrow 'Age of tree'. But, how much and what information is available to each individual may vary depending on receiver's prior knowledge about the information source. This is so called 'relativization' [11, p.79] of the information content of a signal, which should not be confused with it is being arbitrary.

Information is also measurable as long as the probability P of an event is known. Let s_a be a state of affairs among a few others of a selection process S , then Surprisal $I(s_a)$ - the amount of information generated at S when it happens in s_a can be calculated:

$$I(s_a) = -\log P(s_a) \quad (1)$$

Information is capable of yielding knowledge and knowledge requires truth, information requires it too. This truthfulness is a necessary condition for DOS (declarative, objective and semantic) information [14]. Therefore, mis-information or false-information is not information. More precisely either is not in our nuclear sense of information, because neither of them is true. It could be 'negative information'(ibid.), i.e., no information at all is carried due to the *equivocation* being of the same amount as that of the surprisal, or no information is received purely due to the receiver's mis-

understanding, even though the information is being fully carried).

2.2 Information Content of a Signal

Communication theory measures the amount of information generated by an event or carried by a signal via the range of events and their associated probabilities. However, It ignores questions having to do with the content of signals, i.e., what specific information they carry [11, p.41].

Dretske defines two semantic conditions on information [11, p.64]:

1. If a signal carries the information that s is F , it must be the case that s is F
2. The quantity of information the signal carries about s is (or includes) that quantity generated by s 's being F (and not, say, by s 's being G)

The above conditions are being formulated into the notion of 'informational content of a signal' (ibid, p65);

"A signal r carries the information that s is $F =$ The probability of s 's being F , given r (and k), is 1 (but, given k alone, less than 1)".

The parenthetical k here stands for what the receiver already knows (if anything) about the possibilities that exist at the source.

The information content of signal can be seen as *what-we-can-learn* from a signal as the following passage shows;

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 [11, p.45].

Another way to look at the information content is the situation semantics [9]. The information content of an utterance can be identified as *embedding situation* which consists of meanly the *utterance situation* and the relations to part of the world directly relevant to the discourse [9, p.218]. The utterance situation is the situation or context in which the utterance is made and received. For instance, if I say to my wife, "we have no salt left", the utterance situation is the immediate context in which I speak these words and my wife hears it. In this case, no salt is left. It also includes both of us as participants (for the duration of the utterance), and the salt (cooking or table salt) I am referring to and many more such as time, location, etc.

2.3 Semantic Content of a Signal for a Human Receiver

When a signal is observed by a receiver, it is possible to give rise to the *semantic content* of the signal which reveals what a human receiver has learned from the signal by observing it. Human perception and cognition

are heavily involved in information-handling process. Cognitive states always, either explicitly or implicitly, have a specific propositional content [11, p.154]. *Propositional content* of an utterance (signal in general) is that item of information contained in the utterance in the most intentional manner, as opposed to any other information that the utterance might convey, such as the speaker's ability to speak English [9, p.88]. Explicitly speaking, the propositional content is the claim proposed by a signal. The creator of the signal must possess some sort of cognitive state to be able to purpose this propositional content to the signal. For example, the propositional content of the utterance "I have a toothache" is the claim that I am having a toothache.

The structure's *semantic content* is the unique propositional content exhibiting higher order intentionality [11, p.173].

Higher Order of Intentionality:

- (a) It is nomically or analytically necessary that Fs be G
- (b) S has the content that t is F
- (c) S does not have the content that t is G

To illustrate it in simply form, one might know $t=2$ without realizing that t is square root of 4, although by scientific law 2 is indeed the square root of 4.

It is selectively sensitive to that component of the incoming information that defines the structure's semantic content [ibid, p.180], which can be identified to be *the outmost information shell*, the piece of information in which all other information carried by S is nested (either nomically or analytically). It is the one piece of information carried in *digital* form [ibid, pp.177-178]. S carries the information that t is F in digital form if and only if that is the most specific piece of information about t that S carries.

It should be noted that semantic content is relative to receivers in that different individuals may digitize different propositional content as their semantic content of the signal due to their different background of domain knowledge, purpose and their intentionality.

To apply the above notions to databases, data are created by some human originators, and normally have propositional content. However, data in a machine (non-human and therefore has no cognitive capability) do not have the higher order of intentionality, and therefore data or databases do not have semantic content but only propositional content, which may have many other pieces of information nested in it. These are sources of information for a user of the database to apply a process of digitalisation in order to receive the information that interests them. That is to say, from the semantic content of some intention of the database operator, which is only one piece of information without anything else nested in it, to the information content of the data that is placed into the database by the operator, the number of pieces of information is greatly increased. And because of this inflation of information, users may abstract one piece of

information (the additional one) without paying attention to the one proposed by the database originator. That is how database may provide more information than that which is intended to its users.

When the semantic value of a datum is not part of the information content of the data (which may often be the case), and if the former and nothing else is received, then the receiver of the datum would be misinformed. In such a case, we say that the database where the datum resides does not provide the user with any information. We would not use the term 'misinformation' or 'wrong information' as information has to be contingently true [11] [14].

2.4 Concept and Meaning

Concept and *Meaning* are two notions that are closely related to information and representation. For human agent, information is literally untouchable as one is always already in a system of meaning [19].

Meaning, like belief (and other cognitive states), exhibits a higher-order intentionality than that associated with a structure's (e.g., a signal) informational content [11, p.175]. It is the semantic content of a concept. Meaning should never be confused with information. Information embodied by a signal (linguistic or otherwise) is only incidentally related to the meaning (if any) of that signal

[ibid, p.44]. Meaning is subjective or inter-subjective as it always exists in human's mind. Another distinctive feature is that meaning does not have to meet the truth condition whereas information does. A message "it is raining today" has meaning but carries no information if it is not true (not raining).

There is, of course, a use of the term 'meaning' that is close (if not equivalent) to the ordinary sense of information. For example, smoke means fire, a blown fuse means the circuit has been overloaded. Meaning in this category is often referred to literal or conventional meaning, context-free meaning [8] and natural meaning [13]. There is another type of meaning that is induced or derived by literal or conventional meaning. This kind of meaning is often regarded as implied meaning (intention [19], nonnatural meaning [11] [16] or meaning-in-use [9]).

The process of meaning generation is not a straightforward one. It involves another important term, *concept*. A *concept* is generated within a cognitive state through digitalization of information [11, p.142]. It gives meaning (the semantic content of a concept) to its instances. The production of meaning involves the digitization of analogue information, the creation of a concept, and the instantiation of that concept. The whole process can be illustrated in Fig 1.

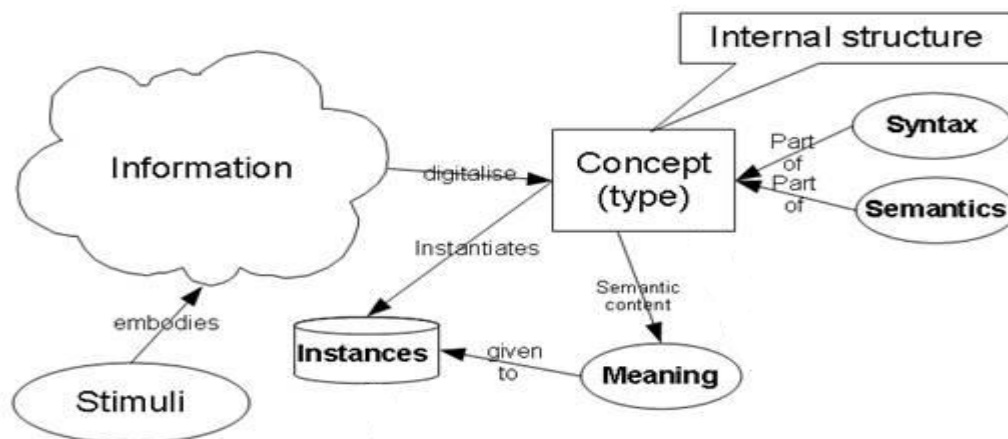


Figure 1: The overall illustration of meaning generation

Many people believe meaning and the semantic content are interchangeable if not equivalent given that both of them possess higher-order intentionality and their close relations to information. They are nevertheless different terms. The semantic content is the digitization of information by a human agent. Therefore, it is contingently true the same as information is. Meaning does not necessarily have to be true. My utterance "there are wolves coming" has meaning but it gives no semantic content to anybody if there is no wolf around. It also generates a wrong belief to people who have little background knowledge that wolves are coming. They react with their belief and their concept on wolf (i.e.,

scary animal, kills the sheep) by running to safety or coming to help. However, after they realise my utterance has not semantic content (no wolves present), a new concept is being developed in their mind that I am a liar. Consequently, my repeat of same sentence later will not give them the same belief and reaction even though this time wolves are indeed coming. But this time the information of wolves are coming is failed to be digitized to be their semantic content.

2.5 Information Content Inclusion Relation

The notion of *Information Content Inclusion Relation* (IIR) was initially proposed by [12] and included in formally published articles such as [26], which studies a special type of relation, namely informational relationship between two events X and Y . Whenever a particular (i.e., instance) of X carries the information that a particular (i.e., instance) of Y exists we say that Y is in the information content of X , which is denoted as $I(X) \ni Y$. It means that a suitable placed observer could learn that Y (particulars of Y) by consulting the particulars of X . The sufficient conditions of $I(X) \ni Y$ are that both of X and Y are events in one probability space and $X \subset Y$, meaning whatever X is true, Y is true.

The strength of IIR lies on its strong reasoning power. Through using IIR rules which were also first proposed by Feng in [12] and then included in [25] among others, hidden information may be derived by finding new IIR from a given set of IIR. There are six inference rules for logically reasoning about information content;

Sum: If $Y = X_1 \cup X_2 \dots \cup X_n$, then $I(X_i) \ni Y$ for $i = 1, \dots, n$

Product: If $X = X_1 \cap X_2 \dots \cap X_n$, $Y = X_i$ for $i = 1, \dots, n$, then $I(X) \ni Y$

Transitivity: If $I(X) \ni Y$, $I(Y) \ni Z$ then $I(X) \ni Z$

Union: If $I(X) \ni Y$, $I(X) \ni Z$ then $I(X) \ni Y \cap Z$

Augmentation: If $W = W_1 \cap W_2 \dots \cap W_n$, Z is the product of a subset of $\{W_1, W_2, \dots, W_n\}$, $I(X) \ni Y$ then $I(W \cap X) \ni Z \cap Y$

Decomposition: If $I(X) \ni Y \cap Z$ then $I(X) \ni Y$, $I(X) \ni Z$

3 The Notion of Representation

It is commonly agreed that representation is a prime notion of information systems because the representational relationship is a fundamental relationship between two states of affairs concerning whether one carries information that the other exists. We argue that although numerous researches have been done toward representation, the content of a representation in particular the semantic content of a representation that connects human designer(s) and user(s) has not been theoretically studied.

3.1 Literature review on Representation

Over past decades, the notion of representation was studied from many different angles. The result shows that representation is highly relevant to presentation of information and information flow. Such a connection has been identified by [2] [3] [9] [23].

According to Devlin, the basic 'item of information' is dubbed 'infons', which are a set of objects with their associated relations. An infon is true within a situation. The common property of situation is its type. Situation

types may be connected, which are captured by the notion of constraints, and a constraint is an informational relationship [9, p.39]. Under the situation theory, representation is modelled from information carrying perspective. Information flow takes place when an agent applies a constraint to a situation. If an individual situation of one situation type is held through 'anchoring' parameters in the situation type, then due to the constraint, an individual situation of the other situation type is also held as a consequence of the former, also through anchoring parameters. What is made true by the latter is said to be in the information content of what is made true of the former. Constraint triggers one situation type to be held as the result of another situation held.

Barwise and Etchemendy [1] assess representation and associated inferences beyond linguistics. The term of 'constraint infon algebra' is being used to analyse inference over information. Also the models of infons are formulated by 'distributed lattice'. Heterogeneous presentations of information content are captured in an independent way while the 'entailment' relations are found in the context of situation theory.

Shimajima [23] studies representation by accounting for the efficacy of two different modes of representation, namely linguistic and graphical representation [ibid, p.vii]. He was the first to identify a number of phenomena, including free rides, over-specificity and self-consistency that are unique to graphical representation. Shimajima's framework captures constraints between the 'source' and 'target' of a representation system by defining 'constraints projection'. This mechanism is later formulated based on the 'information flow channel' to be the 'representation system' [3, lecture 20]. However, Information Flow theory (IF) does not cover the informational relationship between source and target of a representation in great detail. More significantly, the notion of the content of a representation is not fully addressed.

The emergence IF brought some fresh ingredients to the nature of representation.

Its motivating idea is that representation is the result of information flow within a distributed system. An IF based model for representation systems consists of three aspects: an IF channel, *token connections* (i.e., tokens of the core) and associated constraints.

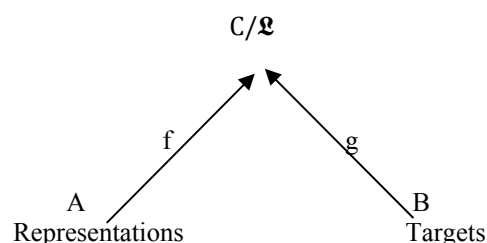


Figure 2: An IF based representation system

This representation system seems working well in circumstances involving cognition. But it still lacks consideration of the content of information flow. Only *normal tokens* are able to support representations if there are constraints on the core that support token connections. However, a token connection being ‘normal’ is neither sufficient nor necessary for making a representation accurate [24].

In terms of how to define it, Shimojima’s definition of representation is seemingly the reflection of its true nature, “Representations are external objects that we use to present information about some other objects on the basis of some fixed semantic rules” [23, p.vi]. However, we argue that representations are not only being used to present information about other objects, more often they are indications of certain situations (events). For example, the name of Human on the note represents the event that Human has been selected among other candidates; a V hand gesture indicates victory of a given situation such as a football game.

Therefore, it is desirable to amend the above definition of representation given by Shimojima as follows:

Representations are external objects that we use to present information about some other objects or situations on the basis of some fixed semantic rules.

The ‘fixed semantic rules’ here are equivalent to *constraints* in IF terms. For example, the traffic lights are the representation of traffic instructions to drivers with constraints such as RED \Rightarrow STOP.

Representations should not be confused with signs. From the above definition, we observe that representations exist in a physical form, contain a significant component - information, the information about another state of affairs. Therefore they are qualified as signs. However, representations will not exist without the involvement of human agents given that they are being used by human originators to serve its users. Signs for example the tree stumps in the forest signifying the age of the tree do not necessarily have cognition involved.

3.2 Quantitative measurement of representation base on ‘information carrying’

We have learned that every *representation system* consists of representations and the represented (target) and there have to be an informational relationship between them (via the core of an information channel). Because there is information flow between two states of affairs, one concerning representations and the other the represented, representational relationship is qualified as Dretske’s [11] ‘information carrying’ in that representations and the represented are the information carrier and the information source respectively. Therefore, how much a target is represented by its representation can be quantitatively measured by using

Dretske’s theory [ibid]. A new term *representation ratio* will be introduced in due course.

- 1) The quantity of information generated in the *represented* (information source) S:

$$I(s_a) = -\log P(s_a), \quad (2)$$

Where $P(s_a)$ is the probability of s_a . This is the *surprisal* of s_a . The weighted mean of surprisals of all random events of S, denoted as $I(S)$, is called the *entropy* of S

$$I(S) = -\sum P(s_i) \log P(s_i), \quad i = 1, \dots, n \quad (3)$$

- 2) The quantity of information generated in *representation* (information carrier) R:

$$I(r_a) = -\log P(r_a), \quad (4)$$

where $P(r_a)$ is the probability of r_a . The entropy of R is denoted as $I(R)$

$$I(R) = -\sum P(r_j) \times \log P(r_j), \quad j = 1, \dots, n \quad (5)$$

- 3) The quantity of information carried from *represented to representation* $I_s(r)$:

$$I_s(r) = I(s) - E(\text{equivocation}) = -\sum P(s_i) \log P(s_i) - (-\log P(s_i/r_j)) \quad (6)$$

- 4) The quantity of information created in *representation* not accounted by the *represented* $I_s(r)$:

$$I_s(r) = I(r) - N(\text{noise}) = -\sum P(r_j) \times \log P(r_j) - (-\log P(r_j/s_i)) \quad (7)$$

- 5) *Representation ratio R*, which we would use to quantify how good a representation system is, is defined in terms of the percentage of information that is transmitted from the *represented to representations*:

$$R = I_s(r)/I(S) \quad (8)$$

Two extreme cases are $R = 1.0$ (*full representation*) and $R = 0$ (*null representation*)

3.3 The Semantic Aspects of Representation

It would appear that the study of the semantics has become an increasingly significant factor in various information system (IS) fields due to its direct bearing to human users. As indicated earlier, the content aspect, especially the semantic content of representation has not been adequately studied by the influential theories of representation.

3.3.1 The information Content of a Representation

As representation is a special type of signs (signals), the information content of a signal is the information content of representation if the sign is a representation. A representation always has a prepositional content that its human originator wants to convey, which is what you could learn from the representation in the simplest sense, and it is part of the information content. Take the same traffic light example, the light is turning into RED has many information contents such as ‘the light is working’, ‘it is in certain size and height’, and many more. However, it crucially includes ‘Stop in front of the line’ as its information content, which is the proposition from the original creators of traffic instruction. We observe that it is this prepositional content that differentiates representations from ordinary signs.

Therefore, the information content of representation can be defined as *what a representation is capable of ‘telling’ us, telling us truly, about the representation per se plus the represented state of affairs.*

We also look at the ‘Information Content Inclusion Relationship (IIR)’ [22] between the represented and the representations in the sense that if $I(X) \ni Y$ (the information content of X includes Y), then X is a full representation of Y (see justification in section 4). We would therefore identify a sufficient condition for X to represent Y from the notion of IIR, namely, (1) both X and Y are events and therefore they could be contingently true or untrue but are neither necessarily true nor necessarily untrue; (2) whenever X is true, Y is always true. That is, in other words, $Y \supset X$.

3.3.2 The Semantic Content of a Representation

In theory, Semantics is the study of meaning [11, p.46] [24]. So far we have learned that representations are external objects that we use to present information about some other objects or situations on the basis of some fixed semantic rules. Representations are supported and sustained by information flow. Also, the semantic content in general is the unique propositional content exhibiting the higher order of intentionality. It is the digitization of analogue information by individual receiver.

The semantic content of a signal can be seen as the receiver’s relativisation of digitisation on this particular signal. There is no fixed semantic content for a signal and a signal does not have semantic content if there is no receiver of it. When a sign is being used by a human agent and therefore becomes a representation, the specific propositional content is embedded into the representation by the human agent, which means it has become a part of the informational content carried from the represented situation (states of affairs). If it is received, this specific propositional content becomes the semantic content of the said representation for this particular receiver.

Therefore there are three elements involved in the ‘representation’ process: the represented situation (say something in the real world), the originator of the representation (a human agent, normally) and the representation *per se*, and the receiver (another human agent, normally) of the representation. The represented situation, for example, Herman is being selected creates normally more than one piece of information due to information nesting. The originator of the representation, for example, the supervisor of the eight employees who were candidates of the selection process, digitalises the information created and decides to have one of them represented and then conveyed, for example, ‘Herman has been chosen’ and nothing else, which involves the higher order of intentionality of the supervisor. This way ‘Herman has been chosen’ now becomes the semantic content of his/her intention. The representation could simply be that Herman’s name appears on a note. The note (i.e., the representation) carries ‘‘Herman has been chosen’ and any information that is nested in it, for example, ‘a man has been chosen’. Finally, the receiver, say the manager, may receive ‘a man has been chosen’ from the note and nothing else, which becomes his/her semantic content of the note – a representation. This involves the higher order of intentionality of the manager.

Thus, the *semantic content of representation* is relative to whom we are talking about – the originator of the representation or the receiver of it. For the former, it is the unique propositional content being selected by its human originator through digitisation of information that is carried from the represented states of affairs, whereas for the latter, it is the information that is received by the human receiver from the representation. Therefore, the semantic content of representation on either account exhibits a higher order of intentionality. The originator of a representation (a linguistic or graphical sign normally) may well use the literal or conventional meaning of the sign to represent the semantic content of his/her intention as the former is fixed and being shared among its users. For example, a blue area in a map has ‘water’ as its conventional meaning.

We re-iterate that the second semantic content of a representation discussed above is directly determined by its users, through their individual digitisation of the representation. We suggest calling it *user’s semantic content* of the representation, and the first semantic content determined by the originator of the representation may be seen as the *intended semantic content* of the representation. These two can be different, and it may accidentally be the same. When this happens, the user has learned what he is supposed to learn from the creator of the representation. Good representation systems are those that enable the best possible match between the two kinds of semantic content of a representation.

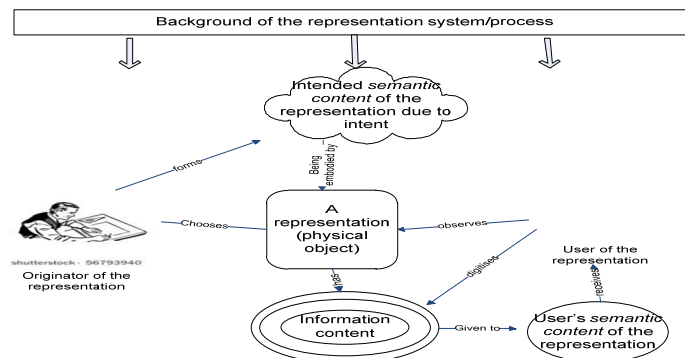


Figure 3: the relationships between representation and its related situations. Note that a different user may digitise a different piece of information content to be his/her semantic content

3 Applying the notion of ‘Representation’ to Connection Traps in a Database

We suggest that databases are a specific type of representation systems within which data values carry information that represents states of affairs in the real world. However, questions such as what kind of a representation system that database is and why and how a database often provides ‘wrong information’ (which is not information at all following [14]) still remains unclear. Moreover, the loss of important information capacity seems one of the main problems in ER [7] modelling. We observe that the notion of ‘representation’ underpins database systems, and that the problem of retrieving information relevant to a need is essentially a problem of understanding the nature of signs [20]. It would seem desirable to apply our representation framework together with the theory of semiotics to tackle some of the fundamental problems of databases.

4.1 Connection Traps in Database Schemata

Connection traps are typical problems in ER modelling of database design. It often happens when two or more relations are joined together to make up of a connection within a schema, which we wish to call ‘schema connection’ in comparison with relations of objects in the real world. Codd [8] who is the founder of the *relational database* concluded that the ‘plurality of joins’ (due to ‘many:1/1:many’ pattern) was the main cause of connection traps. Cardenas [5] believes that connection traps are generated by ‘represent a ternary relation as two binary relations’. Howe [17] suggests that connection traps normally occur due to a misinterpretation of the meaning of certain relationships. The most recent research on connection traps appears to be from Feng [12], who refined the

concept of connection traps and systematically examines the essence of this problem. The result shows that connection traps are caused by the mismatch between the capacity of a path of representing information and the information required to be represented. He further proposed a general solution of connection traps, the separation of *schema connections* that are made possible by a path and *the real world relations* that are required to be represented by the path.

4.2 Our Approach on Connection Traps

Our approach to connection traps may be seen as further development from Feng’s [13] work. We embark on a representational analysis of ER schema, in particular, schema connections. Our mechanism consists of the aforementioned *representational framework* and IIR. It would appear that any kind of connection traps is generated due to two or more schema connections that produce plurality joins [8]. We call a schema connection that possesses no connection traps a *joinable connection*. Any connection trap is a result of the partial representational relationship within a schema connection. A *full representation* between schema connections would be required to make a sound and complete database system. We show our results in the paragraphs that follow.

Definition 1 [3]: If $\alpha \in \text{tok}(A)$ and $b \in \text{tok}(B)$, α is a representation of b if α and b are connected by some $c \in C$. The token α is an accurate representation of b if α and b are connected by some normal token of the core.

Definition 2: A state of affairs r_j is a full representation of a state of affairs s_i if they are connected by an information channel and the equivocation in relation to s_i and r_j is 0 bit.

Definition 3: $I(X) \ni Y$, it means that a suitable placed observer could learn that Y (particulars of Y) by consulting the particulars of X . The sufficient conditions of $I(X) \ni Y$ are that both of X, Y are events and $X \subset Y$, i.e., whatever X is true, Y is true [23].

Proposition 1: if $I(X) \ni Y$, then there must exist an information flow channel between X and Y

Proof

Let x_1 be an instance of event X

$I(X) \ni Y$

$X \subset Y$, Y is also an event (the sufficient condition of $I(X) \ni Y$)

There must be an instance of event Y , say y_1 such that (x_1, y_1) satisfies X and Y . (x_1, y_1) can then be seen as a token of a codomain, say C , which has X and Y among others as its types. Therefore,

$(x_i, y_i) \models X_i, (x_i, y_i) \models Y_i$ premise

Let $f^\wedge(X_i) = X_i, g^\wedge(Y_i) = Y_i$, for $i = 1, \dots, n$

$f^\vee(x_i, y_i) = x_i, g^\vee(x_i, y_i) = y_i$,

for $i = 1, \dots, m$ premise

$f^\vee(x_i, y_i) \models X_i$ if and only if $(x_i, y_i) \models f^\wedge(X_i)$, $g^\vee(x_i, y_i) \models Y_i$ if and only if $(x_i, y_i) \models g^\wedge(Y_i)$, which satisfies the fundamental property of infomorphisms. f and g thus qualify as infomorphism between X, Y and C respectively.

$X \subset Y \implies X \vdash_{XY} Y$

(Comment: by meaning of whenever X is true, Y is true)

$f(X) \vdash_C g(Y)$

(Comment: by the introduction rule of infomorphism)

There is a regular theory $T = \langle C, \vdash_C \rangle$ in C with constraint $f(X) \vdash_C g(Y)$

(Comment: see definition of regular theory [3, p.119])

C is the classification generated by T where its tokens are (x_i, y_i) and types are (X_i, Y_i) .

Therefore, there is information channel between X and Y as they are components of the channel and C is the core classification of such channel with a pair of infomorphism f and g connecting X, Y to C respectively.

Proposition 2: if $I(X) \ni Y$, namely, the information content of X includes Y then X is the full representation of Y

Proof:

$I(X) \ni Y$ Premise 1

There is an information channel connecting X and Y

Proposition 1 2

$X \subset Y$ 1 3

(Comment: the sufficient condition of $I(X) \ni Y$)

$P(Y/X) = 1$ 3 4

Figure 4: Representation channel made up of connection of two schemas

2) $A1$ is full representation of $A2$

$I(A1) \ni A2$

$A1 \subset A2$

Copyright © 2011 MECS

(Comment: the probability of Y under the condition X is 1)

$E(Y/X) = -\log P(Y/X) = -\log 1 = 0$ 4 5

(Comment: the equivocation E in relation to X and Y is 0)

X is full representation of Y (2 and 5) 6

QED

Proposition 3: In ER schema, if there is one to many relationship between Entity A and Entity B , then the many side B is a full representation of the one side A .

Proof

Let A and B be random events Premise 1

(Comment: sufficient condition of ER schema)

Entity A has one to many relationship with Entity B

$I(B) \ni A$ 2 3

(Comment: whenever B is of a certain value say $b1$, A is of a certain value say $a1$. That is to say, that an instance of an entity say X happens to be, say $x1$, may be seen as an event. An entity is therefore not an event. That an entity has a value that cannot be certain beforehand is a random variable as a result of such thing)

B is full representation of A

3 and Proposition 2 4

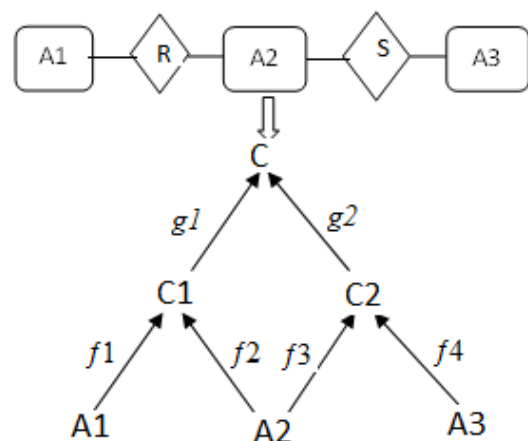
QED

Proposition 4: Given two joint relations $A1, A2$ and $A2, A3$ of ER schema, if $A1$ is full representations of $A2, A2$ is full representation of $A3$. Then $A1$ is full representation of $A3$.

Proof:

There are two ways to justify this proposition.

- Given two relations $A1, A2$ and $A2, A3$, we can build an associated representation channels $C1$ and $C2$, then an integrated channel C that is illustrated in Figure 4.



$A2$ is full representation of $A3$

$I(A2) \ni A3$

$A2 \subset A3$

$A1 \subset A2 \subset A3$

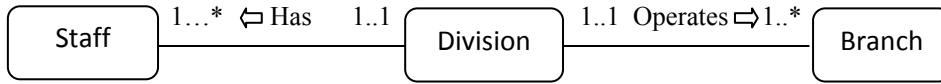
$A1 \subset A3$

$I(A1) \ni A3$

A1 is full representation of A3 *Proposition 2*
 QED

Proposition 5: Connection traps are results of partial representation between schema connections
 We take two typical examples of connection trap, fan trap and chasm trap to justify the above proposition.

Example of fan trap:



Problem:

Division \supset Staff and Division \supset Branch	Premise	1
I(Staff) \ni Division		2
I(Branch) \ni Division		3
I(Staff) \ni Branch	2 and 3	4
Staff is partial (not full) representation of Branch		5

Solution: re-arrange the sequence of entities of schema connections to make it joinable, such as;



Staff \subset Branch \subset Division
 I(Staff) \ni Division (Transitivity rule of IIR)
 Staff is full representation of Division

Example of chasm trap:



Problem: This schema connection is failed to qualify as joinable connection because of the second part of relation, namely Staff Oversees PropertyForRent. There is at least one property that is not overseen by a number of staff and similarly at least one staff oversees none property.

I (Staff) \ni PropertyForRent
 I (Branch) \ni PropertyForRent

Solution:

The solution of chasm trap would be to avoid information lose caused by 0..1 or 0..* relationship. We need to connect Branch and PropertyForRent directly by identifying the missing relationship between them, i.e., Offers relationship. Therefore, the desired schema connection should look like this;



5. Conclusions and Future Work

The focus of this paper is the content aspect of representation which has been heavily ignored by the academic research fields. Our results show that representation is a special kind of signs that is being

used by human agents to convey information about another state of affairs (object or situation). A representational relationship takes place only because representation carries the information that the other (represented) object exists. It is information flow

within a distributed channel that underpins and sustains representation.

Like every ordinary sign (or signal), representations crucially contain information, the information carried from information source as well as those generated in representation. The amount of information contained in representation can be measured as long as the probability of realization of represented can be quantified. We investigate the semantic aspect of representation by revealing that the information content of representation is what you could possibly learn from it and the semantic content of representation is the unique proposition embedded by human originators, the proposition about represented state of affairs. We therefore conclude the semantic content of representation is what a representation is capable of 'telling' us truly about representation per se plus represented state of affairs.

It has been justified that one state of affairs is the full representation of another if the information content of former includes latter. This representational relationship has significance to IS and its applications. Our endeavour shows that databases are actually representation systems in the way that data and representation are both being used by its creators to carry information about other state of affairs in the real world domain. Schema connections are actually representational relations in database. Our example on connection traps has proven that our proposed representation framework can be a promising avenue to solve existing problems in IS.

It has to be noted that our investigation of representation is still in its initiative stage. A more thorough study on its semantic aspects is much needed to make a relatively sound and complete theory of representation. Until then, our representation framework can be widely accepted and utilized to tackle more theoretical problems in IS.

Reference

- [1] Barwise, J. and J. Etchemendy (1990a), Information, Infons, and Inference. In R. Cooper, K. Mukai and J. Perry (Eds.). *Situation Theory and Its Applications*. Stanford, CA: CSLI Publications.
- [2] Barwise, J. and J. Perry (1983), *Situations and Attitudes*, Cambridge, MA: The MIT Press.
- [3] Barwise, J. and Seligman, J., *Information Flow: The Logic of Distributed Systems*, Cambridge University Press, 2008.
- [4] Bateson, G., *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution and Epistemology*, Chicago University Press, 2000.
- [5] Cardenas A.F., *Data Base Management Systems*, 2nd edn, Allyn and Bacon, inc., Boston, 1985.
- [6] Checkland, P. and Scholes, J (1990), "*Soft Systems Methodology in Action*", Wiley Chichester.
- [7] Chen, P. P., 1976, The entity-relationship model - Toward a unified view of data. ACM Trans. On Database Syst. 4, 4 (Dec.), pages 397-434.

- [8] Codd, E.F. (1970). "A Relational Model of Data for Large Shared Data Banks", *Communications of the ACM*, 13(6): 377-387.
- [9] Devlin, K., *Logic and Information*, Cambridge University Press, 1995.
- [10] Douglas Raber, John M. Budd, (2003) "Information as sign: semiotics and information science", *Journal of Documentation*, Vol. 59 Iss: 5, pp.507 – 522
- [11] Dretske, F.I., *Knowledge and the Flow of Information*, Cambridge University Press, 1981.
- [12] Feng J (2002) Minutes of informal Brain Storming session, Friday 17th May, 2002, 16:00 – 18:20 in F219, in which the term 'IIR' was first suggested, University of the West of Scotland, UK.
- [13] Feng, J. and Crowe, M. (1999). "The Notion of 'Classes of a Path' in ER Schemas", in *Proceedings of Third East European Conference on Advances in Databases and Information Systems, ADBIS'99*, Springer.
- [14] Floridi, L., "Is semantic information meaningful data?," *Philosophy and Phenomenological Research*, vol. 70, 2005, pp.351–370.
- [15] Flynn, D., 1998, *Information Systems Requirements: Determination & Analysis* 2nd ed. London, McGraw-Hill
- [16] Grice, H. P., "*The philosophical Review*, Vol. 66, No. 3 (Jul., 1957)".
- [17] Howe, R., David, 'Data Analysis For Data Base Design', Second edition, 1989, Routledge, Chapman and Hall, Inc, New York, ISBN: 0-7131-3688-X.
- [18] Lewis, P.J., "Linking soft systems methodology with data-focused information systems development," *Information Systems Journal*, vol. 3, 1993, pp. 169-186.
- [19] Mingers, J.C., "Information and meaning: foundations for an intersubjective account, *Information Systems Journal*, vol. 5, 1995, pp.285-306.
- [20] Raber, D & Budd, J.M (2003). Information as sign: Semiotics and information science. *Journal of Documentation*, 59(5), 507–522
- [21] Shanks, G., (1999) *Semiotic Approach to Understanding Representation in Information Systems*. In Proceedings of the Information Systems Foundation Workshop (1999).
- [22] Shannon, C. and Weaver, W., *The Mathematical Theory of Communication*, University of Illinois Press, 1949.
- [23] Shimojima, A. (1996). *On the Efficacy of Representation*. PhD Thesis. The Department of Philosophy, Indiana University.
- [24] Stamper, R., "The semiotic framework for information systems research," *Information systems research: Contemporary approaches and emergent traditions*, 1997, pp. 515–527.
- [25] Wilson, B. *Systems: Concepts, Methodologies and Applications* 2nd ed. Chichester, Wiley, 1990.
- [26] Xu, Kaibo and J, Feng, 2009, *Defining the notion of 'Information Content' and reasoning about it in a database*, University of the West of Scotland, Jounel Proceedings of the information Systems Foundations Workshop, Ontology, Semiotics and Practice.
- [27] Wang, Y. and Feng, J. (2009) 'No Representation without Information Flow'- Measuring Efficacy and Efficiency of Representation: An Information Theoretic Approach. WSEAS Transaction on Computers, Volume 8 Issue 3, March 2009.

Author Bibliographies

Lin Liu is a PhD research student in the School of Computing at University of the West of Scotland (UWS). He received his B.Sc. in Mechanical Engineering from the Jilin University, China and his M.Sc. in Information Technology from University of the West of Scotland. His research interests include the Semantic Information Theories and systems, database theory and systems and semantic web technologies.



Junkang Feng BSc, MPhil, PhD was born in Shanghai China and studied at the Shanghai High School in Shanghai and then graduated from the Institute of Military Engineering of the People's Liberation Army (PLA), the Chinese armed forces, majoring in Guided Missiles Engineering. He received his MPhil from the University of Portsmouth, UK and PhD from the University of the West of Scotland (the UWS), UK in Information Systems and Computer Science respectively. He worked as a Research Associate in the Department of Computer Science at the University of Manchester, UK before becoming a Lecturer and then Senior Lecturer at the UWS. He is also a visiting professor of Donghua University and Beijing Union University in China. He currently leads the Database and Semantic Web Research Group of the UWS. His interests include qualitative information and information flow theories, database theory, semantic Web and distributed information systems.