Journal of International Technology and Information Management

Volume 29 | Issue 2 Article 1

2020

The Concept of Representation Capability of Databases and its Application in IS Development

Lishuan Qin
University of Beijing Union University, China, lishuan.qin@buu.edu.cn

Junkang Feng
University of the West of Scotland, UK, junkang.feng@uws.ac.uk

The Concept of Representation Capability of Databases and its Application in IS Development

Lishuan Qin
(Business College of Beijing Union University, China, lishuan.qin@buu.edu.cn)

Junkang Feng
(correspondence author)
(University of the West of Scotland, UK, junkang.feng@uws.ac.uk)

ABSTRACT

The representation capability of an information system in general and a database in particular seems an important and yet elusive concept, which is concerned with, in our view, how a database ever becomes capable of representing real-world objects accurately or otherwise. To explore how to approach and then define this concept, we explore what is meant and required by the statement that a database connection (i.e., a connection between database constructs such as entities in an Entity-relationship (ER) diagram and relations in a relational schema that are made available by a database) refers to, represents and accurately represents a real-world relation respectively. This approach is proven to be insightful and effective. We also find a sufficient and necessary condition for a database connection to be able to accurately represent a real-world relation, which is that the information content of the database connection includes the real-world relation. All these make the concept of representation capability of a database approachable and definable. Furthermore, another different and yet related concept, namely the representation capacity of a database, can also be defined based on the representation capability of a database, which is 'all the real-world relations that can be represented by the constructs that are made possible and available by the database'. Our theoretical work draws on semiotics, the semantic theory of information presented by Dretske and the information channel theory by Barwise and Seligman, and our practical work involves an information system's development.

Keywords: Representation Capability, Database modelling, Database theory, Information content, Information systems

.

INTRODUCTION

The motivation for this work is to explore what enables and is required for a database to represent real-world objects accurately or otherwise, in other words, how a database becomes capable of representing real-world objects accurately or otherwise and thus the *representation capability* of databases. Gregor's paper (2006) in MIS Quarterly says that 'Calls continue for "good theory"' (Watson, 2001) and 'the development of our "own" theory' (Weber, 2003) and presents the nature of theories in information systems. The questions that arise about the bodies of knowledge or theories encompassed in a discipline fall into a number of inter-related classes, and the first one is 'domain questions' (Gregor, 2006). Such questions are concerned with what phenomena are of interest in the discipline, and what the boundaries of the discipline are. We believe that the representation capability of an information system in general and that of a database in particular should be within the boundaries of the discipline of information systems including databases.

To this end, we explore what may be called the 'representational relationship' between a database connection (i.e., a connection between database constructs such as *entities* in an Entity-Relationship (ER) diagram and relations in a relational schema that are made available by a database) and a real-world relation. We find that such a relationship can be divided into three levels bottom-up, namely a database connection *refers to*, *represents*, and *accurately represents* a real-world relation respectively. The first level 'refers to' may be seen achievable based on *semiotics*. To define the other two levels, we draw on *the semantic theory of information* presented by Dretske (1981) and *the information channel theory* by Barwise and Seligman (1997). We develop our solution through theoretical work that involves afore-mentioned theories and database design methods, and a practical information systems development. We also find a sufficient and necessary condition for a database connection to be able to accurately represent a real-world relation, which is that the information content of the database connection includes the real-world relation. All these constitute a seemingly effective means to approach the important and yet elusive concept of the representation capability of databases.

A SEMIOTIC PERSPECTIVE FOR DATABASES

In order to explore how a database construct becomes capable of representing certain real-world objects, we propose an approach that is based on the ideas of semiotics (Stamper, 1997; Anderson, 1997). Semiotics is the study of signs or the general theory of representation (Siau and Tian, 2009). Semiotics has been used to tackle problems in information systems development. For example, Siau and Tian (ibid.) suggest that the graphical notions (or visual signs) of UML are subjected to the principles of signs, and therefore they use semiotics to study the effectiveness of them. We view a database as a collection of signs, and the real-world objects that a database represents are seen as part of properties of signs. Moreover, Stamper (1997) points out, 'signs on every level depend for the correct formation of signs on the level below.' Therefore, a database can be looked at, at least, on two different levels – syntactic and semantic. The former is concerned with the formal structure of the database, and the latter objects and relationships among them that the signs (i.e., data) and constructs of the signs signify. A database design problem may be viewed as a mismatch between the two levels.

A. DATABASE CONNECTIONS VS. REAL-WORLD RELATIONS

A database is constructed according to its conceptual and logical designs, and for the former, ER diagrams are often used, and for the latter relational schemata if the database is to be implemented with a relational database management system. Both an ER diagram and a relational schema are conceptual models in the sense that they specify the data structure of a database but are free from physical and implementation considerations. Conceptual models are widely used, for example, for a clinical decision support system (CDSS), there is a decision model derived from expert knowledge (Hine, Farion, Michalowski and Wilk 2009). The term 'model' is also used to describe mental structures, for example, staff physicians' and residents' mental decision-making models (ibid.). Hine et al. (ibid.) show that the mismatch between the CDSS' model and the residents' model affects triage decision making in emergency room care.

The term 'schema' is also used in many fields of study, and a schema describes a pattern of thought or behaviour that organizes categories of information and the relationships among them. For example, generic surgical process models are mapped onto workflow nets as work-flow schemata to define the behaviour of a surgical workflow management system SWFMS (Neumuth, Liebmann, Wiedemann and Meixensberger, 2012). Callister (2009) observes that experts organise thinking into schemata or mental constructs to both see and solve problem. Schemata are 'organized representations of things or events that guide a person's thoughts and actions' and 'Schemata are the key to expert problem solving' (ibid.). Callister cited Lippman's statement: 'We do not first see, and then define. We define first, and then see' (1961).

.

Relational databases are organised by following the relational data model defined by Codd in 1970. He used the term 'relation' in its mathematical sense of a *finitary relation*. In mathematics, a finitary relation over sets $X_1, \ldots, X_n$ is a subset of the Cartesian product $X_1 \times \ldots \times X_n$; that is, it is a set of *n*-tuples $(x_1, \ldots, x_n)$ consisting of elements x_i in X_i (ibid.). Typically, the relation describes a possible connection between the elements of an *n*-tuple.

We were intrigued to note that 'relation' plays a pivotal role in sensemaking, which is the process by which people give meaning to their collective experiences. Weick (1995) says that 'sensemaking is making something sensible' (p.15), and it is 'about such things as placement of items into frameworks, comprehending, redressing surprise, constructing meaning, interacting in pursuit of mutual understanding, and patterning' (ibid. p.6). Weick (ibid. p.110) cited Upton's (1961) insight: for one thing to be meaningful, 'you must have three: a thing, a relation, and another thing. The meaning of one of them is determined by your momentary awareness of the other two'. Weick says: 'In this book, our unit of meaning has been cue + relation + frame'. Weick states: 'The substance of sensemaking starts with three elements: a frame, a cue, and a connection', and 'Frames and cues can be thought of as vocabularies in which words that are more abstract (frames) include and point to other less abstract words (cues) that become sensible in the context created by the more inclusive words', and furthermore, 'Meaning within vocabularies is relational. A cue in a frame is what makes sense, not the cue alone or the frame alone' (ibid. p.110). Weick observes: 'Frames tend to be past moments of socialization and cues tend to be present moments of experience', and 'The combination of a past moment + connection + present moment of experience creates a meaningful definition of the present situation'. Thus, we have the conclusion: 'If a person can construct a relation between these two moments, meaning is created' (ibid. p.111).

In addition, we observe that Weick's account of sensemaking also sheds further light on the nature of data modelling. For example, Weick (ibid. pp.107-109) cited Freese (1980, p.28): 'Data are not given by experience, but by the concept of the language used to interpret it' and Starbuck and Milliken (1988, p.51): 'Perceptual frameworks categorize data, assign likelihoods to data, hide data, and fill in missing data'.

It would seem that the idea of data modelling could be used in thematic coding, a form of qualitative analysis that involves recording or identifying passages of text or images that are linked by a common theme or idea allowing you to index the text into categories and therefore establish a "framework of thematic ideas about it" (Gibbs, 2007).

Here codes are themes or ideas such as ‘customer service’ and ‘positive’ that are identified from performing thematic coding on customers’ feedback. These codes would then be put into a code frame, which could be hierarchical or flat (Medelyan, 2019) through which the codes like entities in an ER diagram can be visualised and the relationships between the codes captured.

In the context of conceptual database schemata, two types of connections are in question. The connections between data constructs, such as ‘entity’, that are made possible by the topological structure (i.e., a syntactic level formation of signs) of a conceptual database schema or diagram can be termed ‘database connections’ without considering what in the real-world to which they refer. The connections between real-world objects, which is what we want represented by using ‘database connections’, may be called ‘real-world relations’. They are independent of a modelling mechanism such as ER. For example, it might be a real-world fact that employee e1 belongs to division d1, which would be a ‘real-world relation’. If two entity instances, say node e1 and node d1, are connected by an edge in the instance diagram of an ER diagram such as the lower half of Fig. 2, then there is a database connection between them.

A basic task in database design is to construct a sufficient (minimally sufficient if possible) conceptual database diagram or schema that enables all real-world relations that are required to be represented to be actually represented by database connections that are made possible by the diagram or schema. In order to achieve this, we must understand what is meant by that a database connection represents a real-world relation. This takes a few more notions to define.

B. A DATABASE CONNECTION ‘REFERS TO’ A REAL-WORLD RELATION

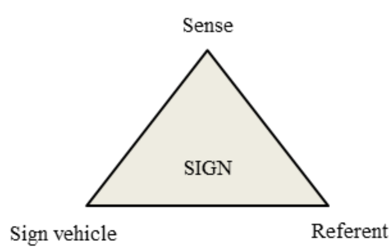

Fig. 1 Peire’s semiotic triad model (Siau and Tian, 2009)

As illustrated in Fig. 1, Peirce’s semiotic triad model shows that the *Representament* (i.e., the form which the sign takes, which is also called ‘sign vehicle’ or ‘signifier’) refers to the *Object* (i.e., the ‘signified’) under the *Interpretant*. The Interpretant is not an interpreter but rather the sense made of the sign. Applying Peirce’s semiotic triad model to databases, a database connection (a

.

’sign vehicle’), say *t*, *refers to* a real world relation, say *s*, if *t* is made up of the entity instances (i.e., nodes in ER instance diagrams in this paper) that refer to the real-world objects involved in *s*, and the link in *t* refers to the link in *s* under the sense-making for database conceptual design.

For example, in Fig. 2 below, node e1 and node d1 form a database connection, and it refers to the real-world fact that employee e1 belongs to division d1. In such a discussion that a database connection refers to a real-world relation, *t* is considered in isolation, i.e., we assume that *t* can be and is already ‘picked up’ from the rest of database connections. The reason for this assumption will be made clear shortly.

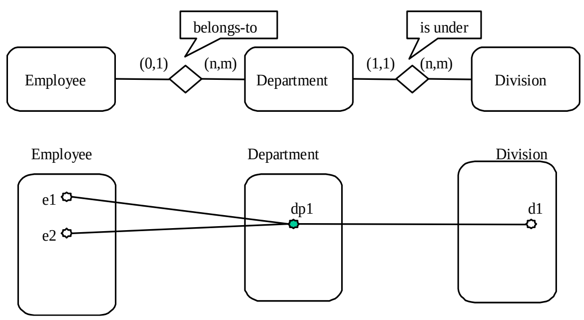

Fig. 2 Database connections shown in an ER diagram

C. RELEVANT AND IRRELEVANT DATABASE CONNECTIONS

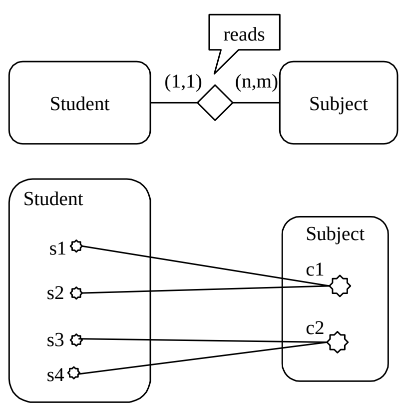

Fig. 3 Irrelevant and relevant schema connections

Due to *nomic structural constraints* (Shimojima, 1996) that a data model has, an instance of a schema normally has extra connections that come inevitably and ‘for free’. For example, in Fig. 2, path (e2, dp1, d1) is such a connection, which is resulted in from the existence of path (e1, dp1, d1) and path (e2, dp1). These unavoidable and ‘free’ connections may have nothing to do with what is supposed to be represented. We call such paths *irrelevant database connections* with regard to a particular set of real-world relations.

More formally, given a collection of real-world relations *S*, a database connection *t* is irrelevant to *S* if it *refers to* no real-world relation in *S*, otherwise *t* is relevant to *S*. Assume that ‘an undergraduate student reads a subject’ is a set of real-world relations. If in Fig. 3, node s1 refers to a postgraduate student, then the connection (s1, c1) is irrelevant to this set of real-world relations.

D. DISTINGUISHABLE DATABASE CONNECTIONS

A database connection must be distinguishable from the rest in order for it to be useful in terms of representing what it is supposed to represent. Let *schema1* be a relational schema or an ER diagram, *t* a database connection made possible by *schema1*, *T* a type of database connections of which *t* is an instance, *S* a set of real-world relations of which *s* an instance; and let *t* refers to *s* and thus it is relevant to *S*. *t* is *distinguishable* regarding *S* if *T* can be explicitly defined by using whatever that is only made available by *schema1*. Moreover, if all irrelevant database connections can be explicitly defined by whatever that is only made available only *schema1*, *T* can also be explicitly defined as a consequence.

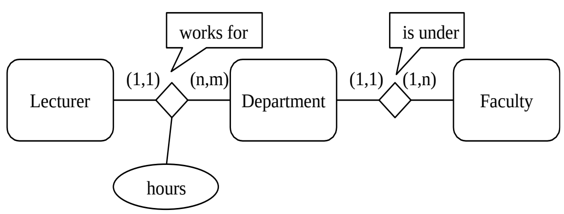

Fig. 4 Relevant database connections that can be explicitly defined

For example, for Fig. 4, assume that only full-time lecturers belong to a faculty, and they belong to the faculty under which the department they work for is. With regard to the real-world relation ‘a lecturer belongs to a faculty’, all database connections *referring to* a part time lecturer and a faculty that are made possible by the path are irrelevant ones. Of all the possible database connections, as long as those that refer to ‘a full time lecturer belongs to a faculty’ can be defined by, say, the post of a lecturer, the hours per week they work, etc, then the relevant database connections are distinguishable. That is, a full-time lecturer might be defined as:

.

Full time lecturer = $\Sigma post_{= FT}$ Lecturer, or
Full time lecturer works for Department = $\Sigma_{hours > 35}$ (Lecturer works for Department)

E. A DATABASE CONNECTION 'REPRESENTS' A REAL-WORLD RELATION

Only when a database connection *refers to* a real-world relation and it is *distinguishable,* can then the database connection be used to indicate that the real-world relation exists. In such a case, we call the former *represents* the latter. More formally, let *schema1* be a relational schema or an ER diagram, *t* a database connection made possible by *schema1*, *S* a set of real-world relations, and *s* an instance of *S*. *t* represents *s* if *t* refers to *s* and *t* is distinguishable regarding *S*.

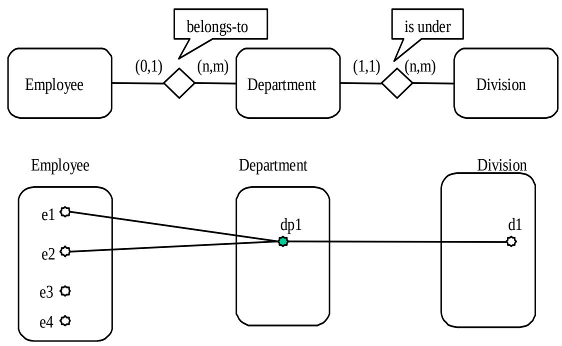

Fig. 5 A database connection is unable to represent a real-world relation due to being indistinguishable

For example, in Fig. 5, which is the same as the one in Fig. 2 where from the discussion earlier database connection (e1, dp1, d1) is relevant while database connection (e2, dp1, d1) is irrelevant. Assume that e1 and e2 do not belong to different proper subsets of the entity, then neither (e1, dp1, d1) nor (e2, dp1, d1) can be explicitly defined by using, for example, relational algebra or SQL. Consequently, the relevant database connection (e1, dp1, d1) cannot be distinguished from the irrelevant database connection (e2, dp1, d1).

It should be noted though that if there is no irrelevant database connection in a path with regard to a type (set) of real-world relations, then the question of whether a database connection is distinguishable does not arise. That is, all database connections represent that set of real-world relations.

The above discussion also shows that a 'representing' database connection must be a 'referring' one first. But the reverse is not true. Fig. 6 in *higraph* (Harel ,1988) illustrates this point, where *t* is a database connection made possible by a database schema, *S* is a real-world relation type, and *s* is an instance of *S*.

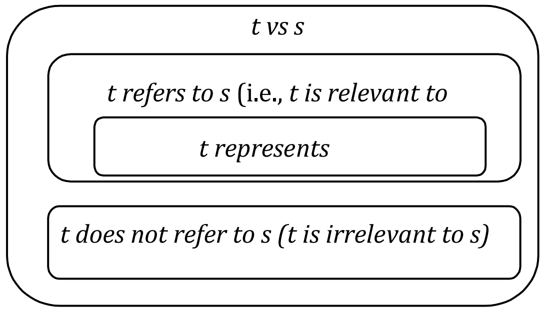

Fig. 6 A database connection refers to or represents a real-world relation

F. PRIMARY MEAMING VS. IMPLIED MEANING OF A PATH

There are certain types of real-world relation(s) that the database connections of a path can always *represent* (also *refer to*, by definition), That is, for such real-world relations, all database connections made possible by the path refer to them, and therefore no irrelevant schema connection is possible. We reveal that such real-world relations are actually the 'primary meaning' of a path. In other words, we define 'primary meaning' of a data construct (Mingers, 1995) in this semiotic way. For a path in an ER diagram, or two or more relations in a relational schema, a database connection made possible by the path or relational join always has a primary meaning. For example, the path in Fig. 7 has the primary meaning that a lecturer delivers a lecture, and a student attends a lecture. These are the real-world relations that the database connections can always *represent*.

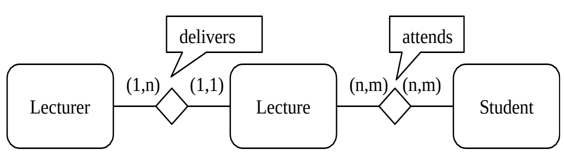

Fig. 7 Primary meaning vs. implied meaning of a path

With certain conditions on both the syntactic level and the semantic level, a database connection may represent a real-world relation that is beyond its primary meaning. For example, the path in Fig. 7 is capable of representing 'a lecturer lectures a student', in addition to the primary meaning that we have just said. All such real-world relations constitute the 'implied meaning' of a path.

.

For the conditions on the semantic level, we look at business rules and the logic of a matter. If a lecturer delivers a lecture, and a student attends the lecture, then the lecturer lectures the student. This is logical. In an organization, there might be a business rule, namely ‘an employee may only work on a project that is controlled by the department to which the employee belongs’. Then from ‘an employee works on a project’ and ‘a project is controlled by one (only one) department’, we get ‘an employee belongs to a department.’

This shows that the meaning of a database construct or a database connection in terms of what real-world objects it can represent is determined with culture as an intervening variable. The afore-mentioned business rules and the logic of a matter are part of the culture or at least under the influence of the culture.

For the conditions on the syntactic level, we look out for the structure of a path. Due to its particular structure, *a path may not be able to provide database connections that refer to a given set of real-world relations, or a path is capable of providing referring database connections, but they are not distinguishable*. We pay attention to the length of the path, the participation constraints of the entities, and so on. When the length of a path is greater than one, we watch out for those situations where the ‘plurality of joins’ (Codd, 1970) may apply. Here we examine the concept of ‘‘plurality of joins’ from the viewpoint that a database connection represents a real-world relation and extend this concept to cover a more general type of database connections. This would hopefully show as an example how we may approach the representation capability of a database.

G. THE NOTION OF ‘PLURALITY OF JOINS’ REVIEWED AND EXTENDED

Codd (1970) puts forward the concept of ‘plurality of joins’ to explain connection traps in a relational schema. For Codd, given two relations R and S, if there are more than one ternary relation U such that $\pi_{12}(U) = R$ and $\pi_{23}(U) = S$, then R and S have the ‘plurality of joins’. For us, more than one U means that more than one set of database connections meet the above criterion (i.e., $\pi_{12}(U) = R$ and $\pi_{23}(U) = S$) and therefore can be established. They are all legitimate syntactically. For example, following Codd (ibid.), we show two joinable relations R and S in Fig. 8, and three different joins of R and S in Fig. 9, Fig. 10, and Fig. 11 respectively below.

R		S	
supplier	part	part	project
1	1	1	1
2	1	1	2
2	2	2	1

Fig. 8 Two joinable relations

R∗S		
supplier	Part	project
1	1	1
1	1	2
2	1	1
2	1	2
2	2	1

Fig. 9 The natural join of R with S

supplier	Part	project
1	1	2
2	1	1
2	2	1

Fig. 10 Another join of R with S

supplier	Part	project
1	1	1
2	1	2
2	2	1

Fig. 11 Yet another join of R with S

.

However, not all joins above represent real-world relations except the 'primary meaning' (what this means was revealed earlier) of the two entities and the relationship between them. Unless a set of real-world relations happens to be matched by the natural join of R and S, at least one database connection does not refer to any of the set of real-world relations. As we said earlier, such a database connection is called an irrelevant database connection. For example, suppose that only (1,1,2), (2,1,1) and (2,2,1) refer to real-world relations, namely 'supplier 1 supplies part 1 to project 2', etc., then (1,1,1) and (2,1,2) are irrelevant database connections. Provided that relevant database connections cannot be explicitly defined (we described this point earlier), a path that is capable of giving rise to 'plurality of joins' will not be able to represent a set of real-world relations that involves all the entities in the path and that is not the primary meaning of the path. For the above example, the result of a join cannot be used to represent the real-world relation that 'a supplier supplies a part to a project'.

This type of situations does not only occur to 'joinable' relations (ibid.). Given two binary relations R and S, as long as π_{21} (R) and S are not functions, that is, they are of many:?/?:many where '?' stands for one or many, for those tuples $\sum_{\pi 2(R) = \pi 1(S)} R$ and $\pi_{23}(U) = \sum_{\pi 1(S)=\pi 2(R)} S$, the same situation occurs. That is, if we let $R' = \sum_{\pi 2(R) = \pi 1(S)} R$, and $S' = \sum_{\pi 1(S)=\pi 2(R)} S$, then R' and S' will be joinable and therefore have the 'plurality of joins'. This would result in the database connections of R and S being unable to represent a set of real-world relations that involves all the entities in the path provided that the real-world relations are not the primary meaning of the path[1]. Thus we propose to extend the idea of 'plurality of joins' to cover any two relations, say R and S, that can have at least one common element in their common column; and to cover a path of length >2 where at least one instance of the entity in the middle of the path can participate in both relationships in the path. That is, given two relations R and S, if there can be more than one ternary relation U such that $\pi_{12}(U) = \sum_{\pi 2(R) = \pi 1(S)} R$ and $\pi_{23}(U) = \sum_{\pi 1(S)=\pi 2(R)} S$, then R and S have 'plurality of joins' (extended from Codd's definition mentioned earlier). Here a U can also be seen as a set of database connections from the 1st column of R to a common element of the common column of R and S, and then to the 2nd column of S. A similar definition of extended 'plurality of joins' for a path in an ER schema can also be formulated.

[1] This conclusion is true under the normal condition, namely the natural join of R and S does not happen to refer to the set of real-word relations and the relevant database connections cannot be explicitly defined.

H. A DATABASE CONNECTION 'ACCURATELY REPRESENTS' A REAL-WORLD RELATION

The above definition of representation does not guarantee that a representation is accurate in the sense that what is represented is actually true. For example, a distinguishable little red circle on a map refers to a school thus it represents the school, however the school is now a club and the map is out of date. Such a representation is not accurate. A database connection may also be an inaccurate representation, when, for example, the database is out of date.

Thus, based on the afore-discussion on what is meant by 'a database connection represents a real-world relation', now we draw upon Barwise and Seliman's formulation of 'representation' (1997, p.235) to define the notion of accurately representation.

For the brevity of the presentation, in the rest of the paper, we use 'path' in a database model to mean any database connection when a database is viewed conceptually as a graph.

The notion of *accurately representation* can be defined as follows: A path say *PathA* in a conceptual database schema or diagram, e.g., an ER diagram, accurately represents a set of real-world relations say *RelA* if for a given instance of a real-world relation, there is at least one distinguishable instance of a path in the database schema or diagram that refers to the given instance of the real-world relation such that the instance of a path is of *PathA* and the instance of a real-world relation is of *RelA*. In other words, the former represents the latter and the latter is indeed of *RelA*. And furthermore, this applies to all possible instances of *RelA*.

The notion of 'accurate representation' may be seen related to some common notions of databases and provide interesting insight about them. Every state of a database is a point on the continuum of the states of the system. The way in which an instance of *PathA* accurately represents an instance of *RelA* as defined here may be used to define the notion of the 'currency of data' in a data storage system in the sense that the data are correct and up to date. Any accurate representation must be current and vice versa. That is to say, a datum's being an accurate representation is equivalent to the datum's being current. Current data as defined here are reliable even though not all reliable data are current, for example, the afore-mentioned little red circle that represents the school that used to exist is not current but still reliable as far as the relevant past states of affairs, namely the situation of the area that the map represents ten years ago, are concerned.

.

Another notion that is related to all these is the 'consistency of data'. In a database, if all data are current, then the data are consistent. This is because real-world relations are always consistent (we may even say that the question of the consistency of real-world relations never arises), and current data are accurate representations of the real-world relations. The converse is not true though, that is, data that are consistent with themselves (even if they are not consistent with the real-world relations) can be obsolete. For example, if the value of a foreign key Department Number of a relational table Employees is 'Dept1' for employee say John, and there is indeed a row in a relational table Departments whose primary key is 'Dept1', then as far as the referential integrity is concerned, these data are consistent. But John has moved to another department, and thus the data are not current.

Thus far, we have identified what constitutes the representation capability of a database construct, which is generalized as a path when a database is viewed conceptually as a graph. The sum of such representation capability is that of the database as a whole. Enabled by the representation capability, all the real-world objects that can be represented by constructs of the database constitute the *representation capacity* of the database.

In the sections that follow we wish to explore the representation capability of a database further by looking at informational relationships between database connections and real-world relations. To this end the notion of the 'information content' of a sign, an event, and in the most general terms, a state of affairs, is relevant. By following Dretske (1981, pp.14-18, 65), this notion is based upon probability and probability distribution.

THE NOTION OF 'INFORMATION CONTENT' OF A STATE OF AFFAIRS

Let us consider the following list:

- Example 1. That there is smoke carries the information that there is a fire.
- Example 2. That he is awarded a grade 'A' for his Programming course contains the information that Jack Brown has gained 80% or above for that course.

Dretske (1981, p.45) defines the nuclear sense of the term 'information content' as follows:

> *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.*

Following Dretske, we take information as in the form of 'de re', rather than 'de dicto', that is, in the form of '*a*'s being *F* carries the information that *b* is *G*'. Dretske (ibid. p.65) establishes the following definition:

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

In this definition, *k* stands for prior knowledge about information source *s*. Dretske's approach, which we will extend for our purposes, is based upon the notion of probability (ibid. pp.14-18), which is concerned with characterizing events, we first give a definition of event:

Definition 1 Let *s* be a selection process under a set *C* of conditions, *O* the set of possible outcomes of s, which are called states, and *E* the power set of *O*, *X* is an event if E϶X and there is a probability of X, i.e., P(X).

The notion of 'probability distribution' applies only within a probability space.

Definition 2 Let *s* be a selection process under a set *C* of conditions, *O* the set of possible outcomes of s, E the power set of O and E϶Xi for *i* = 1,…,n,
Ps is the *probability space* of the events *Xi* for *i* = 1,…,n if Ps = {$P(X_1)$, $P(X_2)$,…, P(Xn)} and ΣP(Xi) = 1.
The information content is concerned with two different levels, namely *tokens* or particulars namely individual things, and their *types* (Barwise and Seliman, 1997, p.69). It is particulars, i.e., individual things in the world that carry information (ibid. p.27). The information that tokens carry is in the form of types (ibid. p.27). Thus, we need a definition for the term 'particulars' of an event.

Definition 3 Let *s* be a selection process under a set *C* of conditions, *X* an event concerning *s, Xi* an instance of *s, Xi* is a *particular* of *X* if *Xi* is in a state **Ω**, written **Ω** = state(*Xi*), and X϶**Ω**.

For example, *s* could be concerned with data values going into an attribute, say, the Emp_Name column of a relational table; *Xi* is a data value in the Emp_Name column at a time *t*, which happens to be 'tony_wu'; the state of *Xi*, i.e., state(*Xi*) = 'a value in Emp_Name column being tony_wu', which is **Ω**; *X* is the disjunction of two states, namely, Ω and say, Γ = 'a value in Emp_Name column being shirley_wu'. Then, *Xi* is a particular of *X*.

.

Given the above concerning the two levels for information content, it would seem appropriate that the above definition of 'the information content of a state of affairs' by Dretske (1981, p. 65) should be modified as follows.

Definition 4 Let *s* be some selection process or mechanism the result of which is reduction of possibilities, and therefore be an information source, and *k* prior knowledge about *s*[2];
Let *r* be an event, and r_i a particular of *r* at time t_i and location l_i;
Let *s*'s being *F* be an event concerning *s*, and s_j some particular of *s*'s being *F* at time t_j and location l_j;
r_i carries the information that there must be some s_j existing at time t_j and location l_j, that is, the state of affairs that *s* is *F* at t_j and l_j, if and only if the conditional probability of *s*'s being *F* given *r* is 1 (and less than 1 given *k* alone).

Definition 5 That a particular r_i carries the information that a particular s_j exists can also be termed that the *information content* of r_i includes s_j, or in other words, s_j is in the information content of r_i.

INFORMATION CONTENT INCLUSION' RELATION (IIR)

Closely following the previous section, given two events, say *X* and *Y*, there might be a special type of relations between them, i.e., 'the particulars of event *Y* are in the information content of the particulars of event *X*'. For brevity, we will also call such a relation 'event *Y* is in the information content of event *X*'. We suggested calling such relations 'information content inclusion relation' (IIR) (Feng, 1998). Interestingly it happens that this term also appears in the literature, for example, in her manuscript, Duží (2001) points out that information content inclusion relations (in relation to attributes) are of partial order.

Definition 6 Let *X* and *Y* be an event respectively, there exists an *information content inclusion relation*, IIR for short, from *X* to *Y*, if every possible particular of *Y* is in the information content of at least one particular of *X*.

An event may have information content inclusion relation (IIR) with more than one other event. Every one of the latter provides the former with its set of particulars, the whole collection of which is 'what a suitably placed observer could learn by consulting' the particulars of the former by following Dretske's definition (1981,

[2] Note that *k* here goes only as far as what counts as a possibility involved in *s*, and it is not concerned with whether an observer is able to learn and actually learns something about *s* by consulting something else such as *r*.

p.45) cited earlier. Therefore, this is the information content of the former. That is to say, the information content of an event is the set of events with which the former has an information content inclusion relation.

Definition 7 Let *X* be an event, the information content of *X*, denoted I(*X*), is the set of events with each of which *X* has an information content inclusion relation.
Therefore, I(*X*) ∍ *Y* is an expression that denotes that event *Y* is in the information content of event *X* through the particulars of event *Y* being in the information content of the particulars of event *X* (For the notion of 'information content', see Definitions 4 and 5 above). For the sake of the completeness of the definition, we allow I(*X*) ∍ *X*, which is a trivial case of I(*X*) ∍ *Y*, when *X* and *Y* are not distinct. Note that in this paper we concern ourselves with the 'information content inclusion' relation as just defined only between events (and their particulars), not any other things. This is because we observe that this event-based approach to looking at databases is helpful.

FURTHER FORMULATING REPRESENTATION CAPABILITY OF DATABASES WITH 'IIR'

Now we explore how the representation capability of a database may be further formulated by means of IIR in order to obtain further insight about this concept.

Proposition 1
Suppose that there is a path *PathA* in a database model/schema and there is a real-world relation *RelA*, the existence of IIR: I(PathA) ∍ RelA is a sufficient and necessary condition for *PathA* to accurately represent *RelA*.

Proof
We prove the 'sufficient' part of the above condition by contradiction. Given I(PathA) ∍ RelA as a premise, then by the definition given above every $r \in RelA$ is in the information content of at least one $p \in PathA$, which means whenever a distinguishable instance *p* of a path happens to be of *PathA*, an instance *r* of a real-world relation is of *RelA*, otherwise *r* may not be of *RelA*. And this applies to every $r \in RelA$. Now let us assume that *PathA* does not accurately represent *RelA*. Then it must be the case that there is at least one $r \in RelA$ such that either no instance *p* of a path such that *p* represents *r* (i.e., either it does not refer to *r* or it does but it is not distinguishable) or *p* represents *r* as being of *RelA*, but in fact *r* is not of *RelA*. This contradicts the premise.

.

We now prove the ‘necessary’ part of the above condition, also by contradiction. Given that *PathA* accurately represents *RelA* as a premise, then by the definition given above, for a given instance of a real-world relation, there is at least one distinguishable instance of a path in the database model/schema that refers to the given instance of real-world relation (i.e., the former represents the latter) such that the instance of a path is of *PathA* and the instance of a real-world relation is of *RelA*. This applies to all possible instances of *RelA*. Now let us assume $\mathrm{I(PathA)} \not\ni \mathrm{RelA}$. Then it must be the case that there is at least one instance of *RelA* such that it is not in the information content of any instance of *PathA*. This means that there must be at least one instance *r* of *RelA* such that there is no any instance *p* of a path such that when *p* is of *PathA* *r* is of *RelA*. This contradicts the premise.

APPLICATION IN AN INFORMATION SYSTEM’S DEVELOPMENT

We applied this concept of ‘representation capability’ in the development of an information system in our college in China to make sure that it can indeed represent what it is designed to represent. This system supports the management of a training centre with over 400 networked computers, and one of the modules of the system is concerned with course/project management. We show a relevant interface of the system below in Fig. 12.

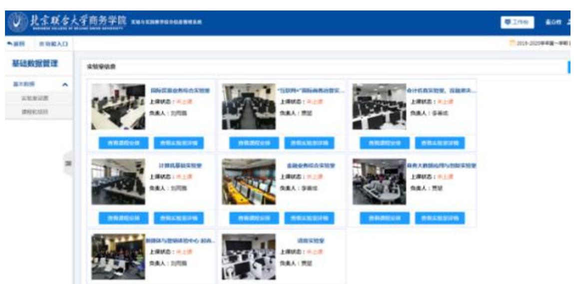

Fig. 12 An information system for a training centre at the Business College, Beijing Union University in China

The conceptual design in the form of an ER diagram of the part of the backend database of the system that is concerned with course management is shown below in Fig. 13.

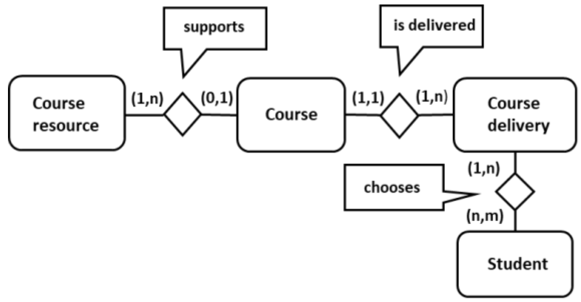

Fig. 13 Part of the backend database for the information system

We now show that the representation capability of the ER diagram enables course management of the course/project management module of the information system. The real-world objects in question are *courses*, *resources* for courses, the *deliveries* of a course and *students* who choose and participate in the delivery of a course. The real-world relations are: ‘a course is supported by various resources such as texts and software’, and ‘a student takes a course’. The ER diagram should be able to accurately represent both.

We justify our design by means of the three levels presented in the paper. First, the development process followed Peirce’s semiotic triad model during which we made sure that when they are considered in isolation all database objects and connections *refer to* the above targeted real-world objects and relations. This is the lowest level, namely ‘referring’ that we have been discussing.

Second, let us show that the database connections enabled by the ER diagram are also distinguishable, and if so, they would be also ‘representing’ the targeted real-world objects and relations.

To this end, we find that the binary relationship ‘supports’ between entity *course resource* and entity *course* would not include any irrelevant database connections regarding the first real-world relation – ‘a course is supported by various resources such as texts and software’, and thus all database connections within this path are distinguishable. The path made up of entities *student*, *course delivery* and *course* does not form a ‘fan structure’, i.e., it is not a structure of ‘many to one’ and then ‘one to many’. Thus, it is not a ‘fan trap’ (Howe, 1989) and no irrelevant database connections with regard to the real-world relation ‘a student takes a course’ are possible. Therefore, all database connections within this path are also distinguishable. This gives us level two, namely ‘representing’.

.

To show that the design is also of 'accurately representing', we only need to make sure that the information content of the first afore-mentioned path includes the first afore-mentioned real-world relation, and that of the second afore-mentioned path the second afore-mentioned real-world relation. Following the definitions in Sections III and IV, it can be seen that with the first path, i.e., the binary relationship 'supports' between entities *course resource* and *course* in place, the probability of the first real-world relation, i.e., 'a course is supported by various resources such as texts and software' is one and otherwise it is not one. Therefore, the latter is in the information content of the former. The same goes with the second path and the second real-world relation. This gives us the highest, i.e., level three, namely 'accurately representing'.

We conclude therefore that the representation capability of the part of the backend database illustrated in Fig.13 enables the functionality of 'course management' of the information system.

CONCLUSIONS

In this paper a seemingly important and yet elusive concept of the *representation capability* of databases has been investigated through theoretical work and practical information systems development. The work presented here draws on semiotics, the semantic theory of information presented by Dretske (1981) and the information channel theory by Barwise and Seligman (1997). It was found that to approach this concept, to explore and identify what is meant and required by that a database connection *refers to*, *represents*, and *accurately represents* a real-world relation respectively is insightful and effective. It was also found that that the information content of a database connection includes a real-world relation is a sufficient and necessary condition for the database connection to be able to accurately represent the real-world relation. All these make the concept of 'the representation capability of a database' approachable and definable.

Furthermore, based on the representation capability of a database, the *representation capacity* of the database can be defined as well, which is all the real-world relations that can be represented by the constructs that are made possible and available by the database.

REFERENCES

Andersen, P. B. (1997). *A Theory of Computer Semiotics*. Cambridge University Press, Cambridge

Barwise, J., and Seligman J. (1997). *Information flow: the logic of distributed systems*. Cambridge University Press, London. ISBN 0-521-58386-1

Callister, P. (2009). Thinking like a Research Expert: Schemata for Teaching Complex Problem-Solving Skills. *Legal Reference Services Quarterly*, Volume 28, 2009 - Issue 1-2: Teaching Legal Research Part I of II pp.31-51

Codd, E. F. (1970). A Relational Model of Data for Large Shared Data Banks. *Com. Of ACM*, Vol. 13 No. 6, 1970, pp.377-387

Dretske, F.I. (1981). *Knowledge and the flow of information*. Blackwell, Oxford

Duží, M. (2001). Logical foundations of conceptual modelling. In: *VŠB-TU Ostrava*, 2001

Feng J. (1998). An Analysis of Some Connection Problems in an EER Schema by using a Set of 'Info Concepts'. *Computing and Information Systems*, University of Paisley, UK, M. Crowe ed., Vol.5 No.2, 1998, pp. 63-72

Freese, L. (1980). The problem of cumulative knowledge. In L. Freese (Ed.), *Theoretical methods in sociology: Seven essays* (pp.13-69). Pittsburgh, PA: University of Pittsburgh Press.

Gibbs, G. R. (2007). 4 Thematic coding and categorizing. *Analyzing Qualitative Data*. London: SAGE Publications, Ltd

Gregor, S. (2006). The Nature of Theory in Information Systems. *MIS Quarterly*, Vol. 30, No.3, 2006, pp. 611-642

Harel, D. (1988). On Visual Formalisms. *Com. of ACM*, Vol.31, No.5, 1988, pp.514-30

.

Hine, M., Farion, K., Michalowski, W., & Wilk, S. (2009). Decision Making By Emergency Room Physicians and Residents: Implications for the Design of Clinical Decision Support Systems. *International Journal of Healthcare Information Systems and Informatics* (IJHISI) 4(2) 2009, DOI: 10.4018/jhisi.2009040102

Howe, D.R. (1989). *Data analysis for data base design*. 2nd edition, Edward Arnold

Lippmann, W. (1961). *Public opinion 81* (7th printing)

Medelyan, A. (2019). Coding Qualitative Data: How to Code Qualitative Research (Updated 2020), https://getthematic.com/insights/coding-qualitative-data/

Mingers, J. (1995). Information and meaning: foundations for an inter-subjective Account. *Info Systems J*, Vol. 5, 1995, pp. 285-306

Neumuth, T. Liebmann1, P., Wiedemann., & PMeixensberger, J. (2012). Surgical Workflow Management Schemata for Cataract Procedures. *Methods inf Med* 2012; 51: pp.371–382, DOI: 10.3414/ME11-01-0093

Shimojima, A. (1996). *On the efficacy of representation*. PhD thesis, Indiana University, Bloomington, IN, USA

Siau, K., & Tian, Y. (2009). A semiotic analysis of unified modelling language graphical notations. *Requirements Eng* (2009) 14:15–26, Springer-Verlag London Limited, DOI 10.1007/s00766-008-0071-72009

Stamper, R. (1997). Organisational semiotics. In: *Information Systems*: *An Emerging Discipline?*, Mingers J. and Stowell, F., ed., McGraw Hill, London, 1997, pp. 267-283

Starbuck, W.H., & Milliken F.J. (1988). Executives' perceptual filters: What they notice and how they make sense. In D.C. Hambrick (Ed.), *The executive effect: Concepts and methods for studying top managers* (pp.35-65). Greenwich, CT: JAI.

Upton, A. (1961). *Design for thinking*. Palo Alto, CA: Pacific

Watson, R. (2001). Research in Information Systems: What We Haven't Learned. *MIS Quarterly*, Vol. 25, No. 4, 2001, pp. v-xv

Weber, R. (2003). Editor's Comments. *MIS Quarterly*, Vol. 27, No.2, 2003, pp. iii-xii

Weick, K. (1995). *Sensemaking in Organizations*. SAGE Publications, Inc, ISBN 13: 9780803971776

.