ISSN 0961-8309

Volume 41(1) Sept. 2020

Xiaoping Shen[1], Lishuan Qin[2] and Junkang Feng[3]

[1]College of Biochemical Engineering, Beijing Union University, Beijing China
xiaoping.shen@buu.ac.uk

[2]Business College, Beijing Union University, Beijing China
lishuan.qin@buu.ac.uk

[3]Database Technology Research Group
School of Computing, Engineering and Physical Sciences
University of the West of Scotland, United Kingdom
junkang.feng@uws.ac.uk
(Correspondence Author)

Exploring the Notion of *Information Content* for Information Systems

SYSTEMIST

Publication of
The UK Systems Society

Published by the UK Systems Society

Registered office: Sidelands, Nutgrove Lane, Chew Magna, BRISTOL, BS40 8PU

Registered Charity, No: 1078782

Editor-in-Chief: *Systemist*
Professor Frank Stowell
University of Portsmouth
Portsmouth
Hampshire PO1 2EG
Email: editor-in-chief@systemist.org.uk

Managing Editor: *Systemist*
Dr Christine Welch
Gatcombe House, Copnor Road
Portsmouth PO3 5EJ
Tel: +44 2392 16 0254
Email: editor@systemist.org.uk

Exploring the Notion of *Information Content* for Information Systems

Abstract

In studies of information systems, the notion of information content is not always explicitly defined. For example, in the field of information integration there are two major approaches namely structural and semantic. However, neither defines the crucial concept 'information'. For another example, in the field of databases, the information content of a database is seen as all possible instances of the database. We believe that this is not satisfactory. In this paper, we explore how we may approach and formulate the notion of information content of an information system. However, there does not seem to be consensus about the ways of thinking about information. We therefore draw on a particular stream in the study of information in the literature, namely qualitative information theories including Dretske's semantic theory of information and Devlin's 'infon' theory, whereby we construct a conceptual framework consisting of information source, bearer and receiver. We use this framework to look at information systems and explore the notion of information content of them. Then we show how this approach may be applied to formulating information content mapping for information integration. Our work presented here is an attempt to find out whether the qualitative information theories that we follow are enlightening and helpful for exploring the notion of information content for information systems, and we find that they are..

Keywords: Qualitative information theories, Information System, Information Content, Infon Theory

Received: 25 May 2020 Accepted: 29 June 2020

1 Introduction

In studies of information systems, the notion of information content is not always explicitly defined. For example, it seems that information integration is one of the most active and exciting fields in information systems research (Bernstein and Haas, 2008). Information sharing from multiple heterogeneous sources is a challenging issue, which ranges from databases, data warehousing, to modern analytic and business intelligence platforms such as Microsoft's Power Business Intelligence (Microsoft Power BI, 2020) and Tableau from Tableau Software (Tableau Software, 2020). In all these fields, the demand for sharing data, information and knowledge increases continuously.

Many approaches have been proposed for achieving those goals in the literature in the past decade. However, the main problems to be faced in integrating information coming from distributed sources are still in the continuing quest, which are related to structural and implementation heterogeneity. These lead to well-known semantic issues, such as semantic heterogeneity, semantic mapping, the application of Semantic Web, etc.

We observe that carrying out semantic studies in these areas may not be sufficient when information becomes the object of the investigation. Furthermore, we observe a dearth of scientific study of the term of 'information' in the database field. The concept of information content has appeared in many published works. However, two main problems appear here. Firstly, clear definitions for information content are not always exposed to the community (Hull, 1984; Miller, Ioannidis and Ramakrishnan, 1993; Blanco, Goni and Illarramendi, 1999). Secondly, some projects use Shannon and Weaver's information theory to measure information content (Arenas and Libkin, 2003; Andritsos, Miller and Tsaparas, 2004). But it is a mathematical model of communication, in which they use probability to define the amount of information that is caused by 'reduction in uncertainty' (Shannon and Weaver, 1949), and therefore it is accepted that the theory is purely quantitative, and covers only the engineering aspect of information creation and transmission (Dretske,1981, p.3).

Motivated by this gap of knowledge identified above we work on an approach for exploring and clarifying notions about information in information systems. We try to answer the question 'what is the information content of an information system?'.

In this paper, we present our findings from investigating the information content of data sources and formulating the notion of information content.

There are various proposals for the formalisation of the concepts of information including qualitative theories of information and quantitative ones (Stanford Encyclopedia of Philosophy, 2018). The former includes the notion of semantic information, which is defined as ‘well-formed, meaningful and truthful data’ (Floridi, 2002). This is close to our everyday naïve notion of information as something that is conveyed by true statements about the world. This notion seems suitable for looking at information systems in general, and it also suits our investigation. Our approach therefore draws on Dretske’s semantic theory of information (1981), Devlin’s ‘infon’ theory (1991) and the information flow channel theory (Barwise and Seligman, 1997), based upon which we construct a conceptual framework consisting of information source, bearer and receiver (SBR for short). We observe that Dretske’s theory starts with a central property of information, namely information reduces uncertainty, and in other words, information is generated through reduction in uncertainty. The amount of information that is generated by some event is the amount of uncertainty associated with the event, and it can be measured by means of entropy: the entropy, H, of a discrete random variable X is a measure of the amount of uncertainty associated with the value of X (Hull, 1984). The main insight of Dretske’s theory for us is the content (not the quantity) of information that a signal carries about an event, and it should be what the receiver of the signal would be interested in. Devlin’s theory gives us ‘infon’ as units of formation, and moreover ‘constraints’, which reveal how a cognitive agent receives information via some information carrier. We will say more on how these theories inform our SBR framework in Section 4.

This paper is organized as follows. We briefly discuss related work found in the literature in Section 2. Then in Section 3 we summarise fundamental concepts of the theories on which we draw, before we proceed to introduce a model for studying information flow in Section 4. In Section 5 we describe our approach to formulating the information content of an ER diagram. In Section 6, we show how our definition of information content may be applied to specifying mappings between conceptual schemata. Finally, we give conclusions and identify future work in Section 7.

2 Literature Review

Before proceeding to our approach, we shall review some related topics, namely, approaches to integrating information and the ‘information content’ of a database. This helps us identify the gap in knowledge in which we aim to fill.

Integrating Information

A general definition of information integration, in a practical sense, is the ability to integrate and retrieve information from heterogeneous and distributed sources to support decision-making and business analysis, etc. Ideally, we expect that information sources talk to each other with a unified interface. A declarative logical representation may need to be used to form the conceptual basis for the basic architecture of information integration systems (Devlin, 1991).

One of the basic requirements for information integration frameworks is to allow the user to specify what information he/she wants rather than how to get it (ibid.). Within a common architecture of information integration, at least one software component (agent) is normally designed for the purpose of managing diverse information sources and describing available information. Some developed systems use mathematical logic as a knowledge representation (KR) technique, such as ‘Infomaster’ (Dimopoulos and Kakas, 2001).

To achieve these goals, two fundamental approaches have emerged in the literature: structural and semantic (Hammer and Mcleod, 1993). A well-known project following the structural approach is the TSIMMIS (‘The Stanford-IBM Manager of Multiple Information Sources’) at the Stanford University (Genesereth, Keller and Dushka, 1997). It uses translator that logically converts the underlying data objects to a common information model. A data description language (model), OEM (Object Exchange Model), is used and the fundamental idea is that all objects, and their subobjects, have labels that describe their meaning. For example, the following object represents weight of Jim of 70kg:

<weight_of_Jim, int, 70>

where the string ‘weight_of_Jim’ is a human-readable label, ‘int’ indicates an integer value, and ‘70’ is the value itself.

Many other projects have been proposed following a semantic approach (Chawathe *et al*., 1994; Arens *et al*., 1993; Bright, Hurson and Pakzad, 1994; Hammer and Mcleod, 1993). MOMIS is one of those approaches to the integration of heterogeneous information. A common data model (ODMI) and language (ODLI) are adopted to describe sharable information. A Description Logics ocdl (object description language with constraints) is used as a kernel language. Information integration in MOMIS is based on schemata and is performed through an extraction and analysis process followed by a unification process. An example of ODLI Class descriptions, showing a relation class is illustrated below:

```
Interface School_Pupil
(source relational School
  extent School_Pupil
  key name)
{attribute string first_name;
  attribute string last_name;
  attribute string id;
  attribute integer year;};
```

The 'Information Content' of a Database

The phrase of information content has appeared in the literature of databases. However, we observe that, firstly, there lacks clear definitions for this important concept. For example, an 'information content preserving' approach to schema transformation is proposed in Rosenthal and Reiner, 1994, in which it says that information content of a schema is defined by the sets of legal states of the database. It would seem that the concept of information content is treated as the same as a well-known notion called information capacity originally introduced by Hull (1984). The concept of information content does not seem to have been formalized and detailed in those works when it is applied onto analyzing schema transformation and integration problems.

Secondly, some research uses Shannon and Weaver's information theory, which is a mathematical model of communication. They use probability to define the amount of information that is caused by 'reduction in uncertainty' (Shannon and Weaver, 1949) as a measure of information rather than to describe the information content. For example, in Andritsos, Miller and Tsaparas, 2004, the authors use techniques of Shannon's theory (Shannon and Weaver, 1949) and define a measure of information content of elements in a database with respect to a set of constraints. Entropy is a measure of the 'uncertainty' of V, a discrete random variable. When its value is high, the certainty with which we can predict the values of the random variable is low. Following the definition of entropy and conditional entropy, the concept called 'mutual information' is introduced as I(V;T). This measure captures the amount of information that the variables convey about each other. The authors use this approach to discovering duplicate, or almost duplicate, tuples and attribute values in a relational instance and to grouping attributes based on the duplication of values. Therefore, it seems that information content is described on the instance level rather on the type level.

In addition, there are some other approaches to representing information content of a database, such as content-free grammar (Levene, 1998), which is used on the information content of individual databases at the instance level, and Automatch (Berlin and Motro, 2002), in which information content of a message is represented by entropy based on Bayesian learning in the field of machine learning.

3 Theories drawn on

Following the previous section, we therefore believe that it is desirable to try and define the notion of information content with sound theories. Let us start with an ontological assumption that information is objective in the sense that 'the generation, transmission, and reception do not require or in any way presuppose interpretive processes. 'In the beginning there was information. The word came later' (Dretske,1981, p.vii).

Dretske's Semantic Theory of Information

The origin of the work that we present here was an attempt to apply the semantic theory of information by Fred Dretske in 1981 onto information systems design. We take Dretske's account of the relationship between information and knowledge to be an important insight, which we intend to use as a way of incorporating epistemological considerations into looking at theories of information.

Dretske (1981, p.45) defines the notion of the 'information content' of a state of affairs 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 focus on claims of the form ‘a’s being F carries the information that b is G’. From the point of view of semiotics, which has been used in developing a science for information systems (Liu, 2000; Stamper, 1997), we say that one signal – a’s being F carries information about a state of affairs – b is G. Relevant to this, Dretske establishes the following definitions:

Definition 1. Let k be prior knowledge about a specific information source, r being F carries the information that s is G if and only if the conditional probability of s being G given that r is F is 1 (and less than 1 given k alone).

Following above definition, we proposed our first basic notion called ‘data bears information’ in Feng, 1999, which is now re-illustrated in Figure 1.

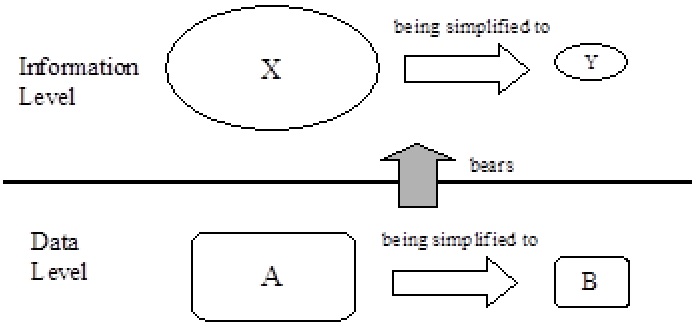

Figure 1 Changes (e.g., simplification) on information level and data level

This diagram indicates that a representation/signal will be considered to represent/carry part of information existing in the real world. When the source of information, namely that part of real world, is changed, a new representation/signal could be used to replace the old one. For example, in the database area, we could use ER diagrams to design a conceptual representation for a university (a part of real world). With the modification made on the information requirements of the university’s information systems, the representation used to bear the information source, namely ER diagrams in this case, would be rearranged accordingly.

Furthermore, for the purpose of making an important difference in the way information can be encoded in a signal or structure, Dretske proposed to use analog vs. digital – in a slightly unorthodox way (1981, pp.136-7).

Definition 2. A signal carries the information that s is F in digital form if and only the signal carries no additional information about s, no information that is not already nested in s's being F.

Definition 3. If the signal does carry additional information about s, information that is not nested in s's being F, the signal carries this information in analog form.

Every signal carries information in both analog and digital form. The most specific piece of information that the signal carries (about s) is the only piece of information it carries (about s) in digital form. All other information (about s) is coded in analog form.

Consider a picture and a statement. Suppose a cup has coffee in it, and a statement 'The cup has coffee in it' is used as a signal to communicate this piece of information that the cup has coffee in it in digital form. No more specific information is supplied about the cup than there is some coffee in the cup. You are not told how much coffee there is in the cup, how large the cup is, and so on. If, on the other hand, a photograph of this scene is shown to you, the information that the cup has coffee in it is conveyed in analog form. The picture tells you that there is some coffee in the cup by telling you, roughly, how much coffee is in the cup, the shape and the size of the cup, and so on.

Therefore, to describe a process in which a piece of information is converted from analog to digital is to describe a process that necessarily involves the loss of information. Information is lost because we pass from a structure of greater information content to one of lesser information content. Analog to digital conversion is a process in which irrelevant pieces of information are pruned away and discarded. In a data integration system, with regard to the specification of the mapping between the global schema and the sources, either global-as-view (GAV) approach or local-as-view (LAV) approach requires an information-preserving mapping. Here the information, we observe, is in digital form. We believe that formulating the information contents of diverse data sources and the mappings between them would be beneficial.

Devlin's 'infon' Theory

Devlin (1991) formulates a mechanism for information to flow based upon the Situation theory (Barwise and Perry, 1983), whereby he explains how a receiver receives information. His logic is based on information rather than truth (such as classical logic) and at the very least capable of handling various kinds of informational transfer (Devlin, 1991, p.12). He aims to define information as a mathematical concept that can form the basis of a scientific study. To this end, he uses Dretske's idea to obtain the kind of precisely defined conceptualization of information, namely information in analog form and digital form.

Devlin defines one of the basic concepts in his theory, namely 'infon'. An infon is a digitalization of information, and a mathematical entity. It corresponds to the way things are in the world as a human agent perceives them. For example, it is perceived by someone that Jim sees a player Alfredo Morelos in a football match between Rangers and Celtic at the Ibrox football stadium in Glasgow on the 24th of July 2019, which can be formalised as an infon.

Definition 4. An infon is denoted as $<P, a_1, \dots, a_n, l, t, i>$, where

1. P is an n-place relation (where n is a natural number);
2. $a_1, \dots, a_n$ are objects appropriate for the respective argument places of P;
3. l is a spatial location;
4. t is a temporal location;
5. i is the polarity of the infon and equal to 0 or 1.

The above expression means that individuals $a_1, \dots, a_n$ have (when i=1) or do not have (when i=0) a relationship P, at spatial location l and temporal location t. For the above example, we would write

<sees, Jim, Alfredo Morelos, Ibrox football stadium in Glasgow, the 24th of July 2019, 1>

in which

1. P is a 2-place relation: 'sees';
2. a1 and an: Jim and Alfredo Morelos, which are appropriate for the argument places of 'sees';

3. l: Ibrox football stadium in Glasgow;
4. t: the 24th of July 2019;
5. i: 1, meaning that Jim sees the football match.

We will use a letter like σ or υ to denote an infon. For example,

$$\sigma = \langle P, a_1, \ldots, a_n, l, t, i \rangle.$$

For the infon above, we may write

σ1 = <sees, Jim, Alfredo Morelos, Ibrox football stadium in Glasgow, the 24th of July 2019, 1>

An infon may include parameters, and if so, such an infon is called a parametric infon. In a parametric infon, there is at least one occurrence of a parameter, and a parameter refers to an arbitrary individual of some type and is a 'representative' for an object. For example, if we replace 'the 24th of July 2019' in infon σ1 with a parameter t, then we would have a parametric infon:

σ2 = <sees, Jim, Alfredo Morelos, Ibrox football stadium in Glasgow, t, 1>

Infons, which is short for parametric infons, are the basic 'informational units' (Devlin, 1991, p.97). Assigning value to parameter is a mechanism called anchoring function. Parameter-free infons (or 'states of affairs') are the basic items of information.

An infon is only true in a certain context, a real situation. A real situation is just a structured part of reality individuated by a human agent, for example, last night's dinner party, a football match, a company, etc. Let *s* be a situation, and given an infon σ, we shall need to consider whether it is a fact that σ is 'made true by' *s*. For example, the football match between Rangers and Celtic on the 24th of July 2019 is a real situation and it makes the above infon σ1 true. If we denote this real situation with s_1, then we would write $s_1 \models \sigma 1$ to means that σ1 is true in the context of s_1.

Definition 5. Given an infon, σ, it is a fact of the world that the relation s ⊨ σ either holds or does not hold. If I is a set of infons and s is a situation, s ⊨ I if s ⊨ σ for every infon σ in I.

For example, let σ3 = <sees, Jim, Jermain Colin Defoe, Ibrox football stadium in Glasgow, the 24th of July 2019, 1>, then s_1 ⊨ I where I = {σ1, σ3}.

Definition 6. An abstract situation consists of a set of infons. It is a mathematical construct built out of relations, individuals, and locations.

Situations (either real or abstract) may be static, involving either just one spatial location or a number of contemporary spatial locations, or they may be dynamic, possible spread over a time-sequence of locations. An example of an abstract situation would be:

s = {<lecturing, Jim, l, t, 1>, <listening-to, Malcolm, Jim, l, t, 0>}

in which two infons constitute an abstract situation s, and it defines a situation where Jim is lecturing but Malcolm is not there.

Information flow is made possible by a network of abstract linkages (called constraints, and more on this shortly) between higher-order uniformities known as types. The constraints that link types are themselves uniformities, which are regularities in the world that an agent can either cognitively individuate or else behaviourally discriminate. In Devlin, 1991, pp.51-2, a set of basic types are:

TIM: the type of a temporal location, e.g., the 24th of July 2019
LOC: the type of a spatial location, e.g., Ibrox football stadium in Glasgow
IND: the type of an individual, e.g., Jermain Colin Defoe
RELn: the type of an n-place relation, e.g., sees
SIT: the type of a situation, e.g., a football match
INF: the type of an infon, e.g., σ3 above
TYP: the type of a type, which is the type for every type. For example, LOC is of type TYP
PAR: the type of a parameter, e.g., t in σ3 above is a parameter
POL: the type of a polarity (i.e., the ‘truth values’ 0 and 1).

Definition 7. Given a SIT-parameter, ċ, and a set, I, of infons, there is a corresponding situation-type

[ċ | ċ⊨ I]

the type of situation in which the conditions in I obtain.

For example,

[SIT1 | SIT1 ⊨ <running, ė, LOC1, TIM1, 1>]

(where ė is a parameter for a person) denotes the type of situation in which someone is running at some location and at some time.

As said earlier, information flow is made possible by a network of abstract linkages (called constraints) between situation types. For example, we may use a constraint to formulate that wherever there is smoke there is fire like this (following Devlin, 1991, p.95):

Let S0 be a situation type of 'there is smoke': S0 = [SIT0 | SIT0 ⊨ <smoke-present, LOC1, TIM1, 1>], and let S1 be a situation type of 'there is fire': S1 = [SIT1 | SIT1 ⊨ <fire-present, LOC1, TIM1, 1>], then we have the constraint:

S0 → S1

Definition 8. Each relation has a collection of argument roles – 'slots' into which appropriate objects or parameters can be filled.

For example, we have an utterance of

Jim sold a car.

It includes three uniformities: the individual Jim, a certain car, and the relation of selling. To denote argument roles in this case, we write it like this:

<takes | seller: P, object-sold: K>

where P is the type of all individuals that are capable of selling things and K is the type of all sellable objects. Some other argument roles maybe left unfilled, such as the location at which the selling takes place.

4 The SBR Framework

In database development, ER diagrams are often used as the conceptual design of a database. It is not trivial to construct an ER diagram for a real-word database development project. With an informational perspective, an ER diagram is an information-bearing signal, which carries information about part of the real world. Human agents receive information about the part of the real world by consulting the ER diagram. These appear to be a coherent whole, and therefore, we believe that information systems in general, and the notion of information content of information systems in particular, could be or even should be approached with a holistic approach of systems (Stowell and Welch, 2012, p.4). Stowell and Welch (ibid. p.233) defines systems as follows:

A system is a set of interconnected parts that, for us, seem to form a meaningful whole which we can name and draw a boundary around it – it is something which, taken together, has meaning for us.

Maturana and Varela (1992, p.74) say that ‘Organization denotes those relations that must exist between the components of a system for it to be a member of a specific class’.

For information systems, we can see three parts, namely some information Source, at least one information Bearer and some information Receiver (SBR for short), which are interconnected through the informational relationship between them – some information flows from one part to another. There parts constitute a system, which we call the SBR framework as illustrated in Figure 2.

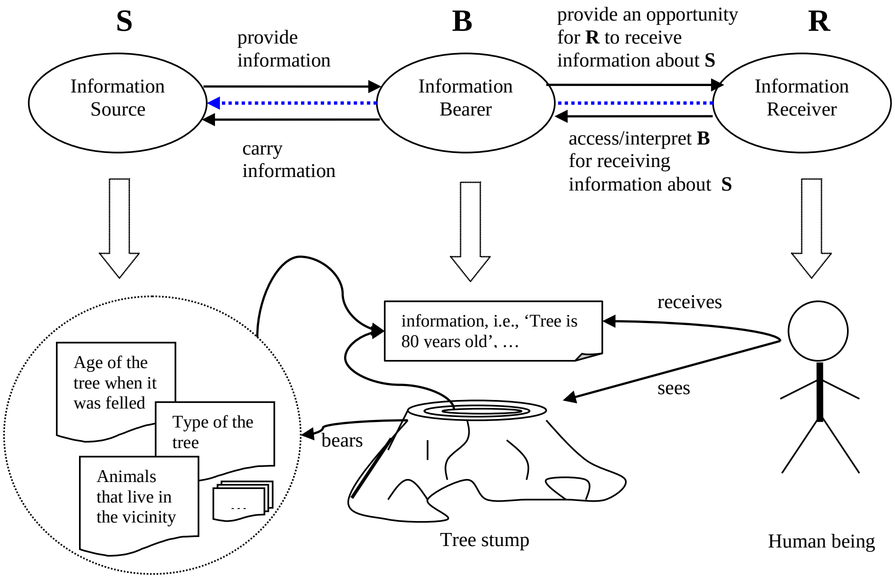

Figure 2 The SBR Framework

As illustrated in Figure 2, some information is created due to reduction in uncertainty following Shannon (Shannon and Weaver, 1949) and Dretske (1981, p.4), for example, the tree happens to be 80 years old, rather than that it is possible that the tree is 40 years old or 80 years old among many other possibilities, so the tree is an information source. This information can be carried by an information bearer, the rings on the tree stump in this case, due to the relationship between the number of rings and the age of the tree, and such a relationship is based upon some 'nomic dependencies' (Dretske,1981, pp.74-5). An information bearer provides an opportunity for an information receiver, for example, a human agent, to receive information about the information source. By consulting an information bearer, an information receiver can acquire information (illustrated by dotted line in Figure 2) if the receiver is aware of and attuned to some constraints (Devlin, 1991, p.15), which formulates the dependency and therefore the informational relationship between the bearer and the source. For this example, the constraint would be a linkage between a situation type concerning the number of rings on a tree stump and a situation type concerning the age of the tree. That is, without using the infon notation,

$$\text{'Number of rings' type} \rightarrow \text{'Age of tree' type}$$

In the field of databases, the part of the real world with which the database is concerned is the information source as reduction in uncertainty takes place in it, for example, a university. The academic programmes and courses in the university happen to be what they are (i.e., they are contingently true) can be seen as some reduction in uncertainty from many possibilities. The ER diagram for the database would be an information bearer as the entities and relationships in the diagram are capable of telling us truly what the academic programmes and courses happen to be. The users of the ER diagram, be it a database developer or a database user, would be the information receiver as they would learn something true about the academic programmes and courses by reading the ER diagram. This is achieved by the user being aware of and attuned to some constraints that link the constructs of the ER diagram and objects in the university, for example,

‘The Name of an Entity Programme’ type → ‘An Academic Programme’ type

We believe that the SBR framework sheds some light on the nature of information systems development and informs the database development process. We start this process with information requirements collection and analysis. This is a matter of identifying and formulating the information that the information receiver R is to receive. Information requirements are normally concerned with a specific domain of the real world, and it is necessary to develop some appreciation of the domain. For this, many writers advocate a soft systems approach, such as (Checkland, 1981; Checkland and Holwell, 1998; Stowell and West, 1994). We combine this approach and semantic theories of information to have elaborated the ‘conceptual model’ for a ‘notional system’ by means of identifying and formulating relevant ‘situation types’ in Feng, 1999. This is an investigation of the information source S of the SBR. Then the information requirements are subsequently satisfied by the ‘informational units’ in the ‘data storage situation type’ and the ‘basis situation type’, and thus we specify the B of the SBR. For details of these topics, interested readers are referred to that paper (ibid.).

The theories described in the previous section, namely Dretske’s Semantic Theory of Information (1981) and Devlin’s ‘infon’ Theory (1991) inform the construction of the SBR framework, especially the following elements, which we summarise here. Dretske takes Shannon’s (Shannon and Weaver, 1949) insight that information is created due to reduction in uncertainty, thus we have the notion of information source S. It is somewhere in which there is uncertainty and then somehow the uncertainty is reduced. Two information sources may be somehow linked in that the reduction in uncertainty of one is affected by the other. Dretske formalises this as the change of probability distribution of one information sources due the existence of the other, and this is the essence of so called the informational relationship between two information sources. A special case of the informational relationship is that the reduction in uncertainty at one information source is accounted for by that at another information source, and therefore the former carries the information about the latter. We call the former an information bearer of the latter. Thus, we have the notion of information bearer B. Devlin's concept of 'constraints' helps us formulate the relationship between S and B. Furthermore, Devlin's idea that an agent is aware of and attuned to a relevant constraint whereby to receive information about S through observing B enables us to have the information receiver R, and how R links with S through B. Moreover, B carries information about S in an analogue form, and R receives information about S by consulting B through a process of digitalisation through which much information is lost. In addition, a general assumption under which the SBR framework is constructed is that the existence of information is independent of its interpreters or receivers (Dretske,1981, p.vii). Once constructed, the SBR framework enables us to look at information flow and information systems from an information theoretical perspective.

In this paper, we concentrate on exploring the notion of information content of information systems, and consider issues on how to formulate, map, and integrate information content of conceptual schemata to show how our approach may be applied. In a case of information integration, we would normally face a collection of heterogeneous data sources, the conceptualisation of which can be in the form of ER diagrams. Therefore, in a task of information integration, some agents (i.e., knowledge engineers, database designers, system users, etc) will be involved to analyze those information bearers (probably with some knowledge of the domain or the help from some design documents) in order to generate mappings and build a global schema if required. We will show our work thus far on this problem in section 6.

The SBR framework enables us to approach the notion of information content of information systems, which would therefore be 'all that the system can tell us truly about the real world, that is, all that, which is contingently true, a suitably placed observer could learn about something in the real world by consulting the system'.

Furthermore, what the system can tell is only contingently true and not necessarily true. Mathematically, the probability of it is between 0 and 1. This follows the well-known theory by Shannon (Shannon and Weaver, 1949).

We believe that our approach is not unlike some known conceptualization for information systems, such as the 'infological equation' formulated by Börje Langefors, which describes the difference between data and information as follows:

$$I = i(D, S, t)$$

This is a mathematical expression for the observation, that the information 'I' communicated by a set of data to humans in an information system is a function 'i' of the data 'D', the semantic background 'S', and the time interval 't' of the communication (Langefors, 1973). Here the 'D' corresponds to our 'B' the information bearer, the 'S' to the background of an instance of the SBR taking place, and the function 'i' may be seen achieved by the cognitive agent's being aware of and attuned to some constraint that links the 'D' and the 'I' in this the 'infological equation'. Therefore, our framework seems to fit with this equation, even though it is not informed directly by the equation.

Another issue related to our discussion is meaning. We agree with Dretske that the confusion of information with meaning is not helpful. 'Meaning, and the constellation of mental attitudes that exabit it, are manufactured products. The raw material is information.' (Dretske,1981, p.vii). Here we show how our framework may help look at how meaning is produced. 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' (Weick, 1995, p.111). Our SBR framework may help shed some light on this process. We may take Weick's frames as patterns of past moments of socialization and they correspond to the 'constraints' in our SBR, and the cues that are present moments of experience are some 'real situation' in which the parameters of the situation types in the constraint are 'anchored' to individual objects, and thus certain information in the form of parameter-free infons is generated and can be obtained by a human agent. This information helps the agent to make sense of the present situation, and therefore meaning is created for the agent.

Mingers (1995) suggests that meaning has three levels. The first level is the primary meaning, which is the semantic content of a sign or a linguistic message, that is, it is the digitalized information without its analogue nesting, and it is obtained by a human agent through the digitalization of the information that is carried by a signal. Our SBR framework can be applied here too. Our B would the sign or the linguistic message, then its semantic content could be formulated as some situation type, which would be part of our S. For example, the utterance of ‘Jim sold a car’ would have the semantic content S1 = [SIT1 | SIT1 ⊨ <sold, Jim, Car, t1, l1, 1>]. A listener who hears this utterance would be our R, who would obtain the information by digitalizing the information that is carried by the utterance.

An information system may be seen having two parts in terms of what carries information: one is the persistent data storage such a database, and the other anything that holds nonpersistent data such as variables of various types, parameters in stored procedures and functions, views and external tables. The information content of either can be formulated by using various means, one of which is situation types that enable certain parametric infons to hold. In the section that follows, we show how this can be done for the conceptual model, in the form of ER diagrams, of a database.

5 The Information Content of an ER Diagram

Following Dretske and Devlin and under our SBR framework, we now propose an approach to defining and formulating the information content of an ER diagram. Our interest here is with the information that agents can pick up about situations in the world. Devlin takes an assumption that every n-place relation for an infon has a fixed number of argument places that cover all the possible arguments the relation has (in any usage), including location and time arguments where appropriate (Devlin, 1991, p.115). This suits our approach because the data sources (databases) that we investigate have a finite number of data constructs, namely, entity types, relationship types, attributes, and data constraints.

Intuitively, the information content of an ER diagram consists of items of information about a specific information source (e.g., a university), as opposed to any other information that the diagram might convey, such as the actual size of a symbol in the diagram and the meaning of the language used for annotating the diagram. Furthermore, in formulating the information content of an ER, we assume that what the ER diagram tells us, i.e., the propositional meaning, namely the semantic content of the ER diagram, is contingently true. We believe that this is a common tacit assumption in database design. The propositional meaning of the ER diagram is therefore the core of the information content of the ER diagram. The whole information content of the ER diagram would be all the contingently true states of affairs that are implied by the core. Due to 'information nesting' as defined by Dretske: 'The information that t is G is nested in s's being F = s's being F carries the information that t is G' (1981, p.71), the whole information content of an ER diagram normally cannot be exhaustively identified. There is no need to do so either because as long as the information that is required is within its information content, the ER diagram is effective. In the rest of the paper, we concentrate on the formulation of the core of the information content of ER diagrams.

Definition 9. Given S - an ER diagram, Given *S* - an ER diagram, we define $e \in IND$, $a \in IND$, and $r \in REL^n$, if $e \in E$, $a \in A$, and $r \in R$
where
1. IND is the type of an individual
2. RELn is the type of an n-place relation
3. E is a complete set of entity types and e is an entity
4. A is a complete set of attribute types and a is an attribute type
5. R is a complete set of relationship types and r is a relationship

All these can appear in S.

Definition 10. Given S - an ER diagram, and k - what agents know about the information source (a domain of application), l - a parameter for the type (LOC) of a spatial location (a database), t - a parameter for the type (TIM) of a temporal location (it is the time when the database, say D, is looked at and S is analysed), the information content of S, is a tuple I = (I(E), I(R), I(E→R), IND, RELn, f, LOC, TIM) where

1. *I*(E) denotes the core of the information content of a finite set of *entity* types in *S* and

$I(_\mathrm{E}) = [s_1 \mid s_1 \vDash (\forall \dot{e} \in E)$ <Having_attributes, $\dot{e}$, $a_1, a_2, \ldots, a_n$, l, t, 1>]

where s_1 is a parameter for a situation of the basic type *SIT*, $\dot{e}$ is a parameter for an entity, 'Having_attributes' is an *n*-place relation of the type REL^n, and $a_1, a_2, \ldots, a_n \in A$.

The above expression means that the information content of the entities in the ER diagram *S* is now formulated as a 'situation type' (Devlin, 1991, p.59) in which there is a parametric infon <Having_attributes, $\dot{e}$, $a_1, a_2, \ldots, a_n$, l, t, 1>], and furthermore, for any entity instance the conditions for the infon obtain, which then becomes a parameter-free infon (i.e., an item of information). We use this expression to formulate the intuitive understanding that the set of entity types in an ER diagram captures all entity instances that satisfy what the ER diagram specifies such as the attributes that an entity has and the relationships with other entities that the entity holds. Notice that this idea applies to all the definitions concerning information content in the rest of the paper.

2. $I(_\mathrm{R})$ denotes the core of the information content of a finite set of *relationship* types in *S* and

$$I(_\mathrm{R}) = [s_2 \mid s_2 \vDash (\forall \dot{c} \in R) \text{<Having\_attributes}, \dot{c}, a_1, a_2, \ldots, a_n, \text{l, t, 1>}]$$

where s_2 is a parameter for a situation of the basic type *SIT*, $\dot{c}$ is a parameter for a relationship, 'Having_attributes' is an *n*-place relation of the type REL^n, and $a_1, a_2, \ldots, a_n \in A$.

3. $I(_{\mathrm{E}\to\mathrm{R}})$ denotes the core of the information content of a finite set of combinations of a relationship type and related entities in *S* and

$$I(_{\mathrm{E}\to\mathrm{R}}) = [s_3 \mid s_3 \vDash (\exists \dot{c} \in R) \text{<}\dot{c}, e_1, e_2, \ldots, e_n, \text{l, t, 1>}]$$

where s_3 is a parameter for a situation of the basic type *SIT*, $\dot{c}$ is a parameter for some relationship, and $e_1, e_2, \ldots, e_n \in E$.

4. *f* is the anchoring function, which assigns to each relation its values for the parameters.

Distinguished from those approaches discussed in Section 2, the concepts that are used in our expressions are not a syntactic or semantic object but are informational objects. They refer to part of reality, a real situation.

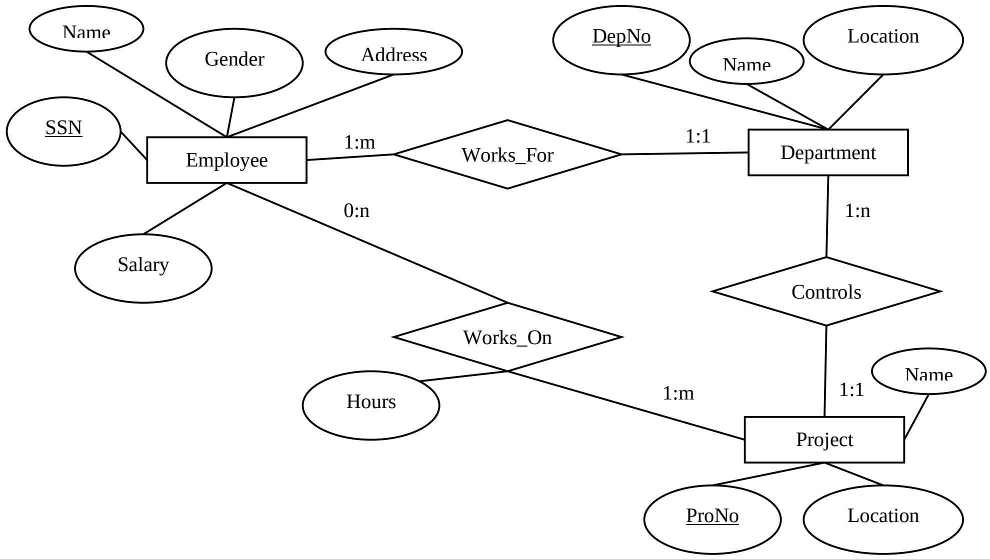

Figure 3 Example of an ER Diagram *S*

We shall now use an example to illustrate Definition 10. Figure 3 shows a simple ER diagram S. We assume that it was designed for some organization's database D at some time T in the past. The core of the information content of S includes:

I(Employee) = [s1 | s1 ⊨ <Having_attributes, Employee, SSN, Name, Gender, Address, Salary, D, T, 1>];
I(Works_On) = [s2 | s2 ⊨ <Having_attributes, Employee, Project, Works_On, Hours, D, T, 1>];
I(E→R) = [s3 | s3 ⊨ <Controls, Department, Project, D, T, 1>];
…

If I(Employee) is anchored in a real situation to a person named 'John' and all those known values of his attributes, we will have a parameter free infon like this:

<Having_attributes, Employee_JS012, JS012, John, Male, 25 High St Glasgow, £35000, D, T, 1>

where 'Employee_JS012' is a combination of the entity type's name and the value of its primary key.

As discussed in section 3, an infon could be parameter-free or parametric, in the above we use these two constructs to capture and formulate the information content of a database on the type level or the instance level. Furthermore, the fact that a particular employee works on a certain project is captured by the use of the same parameter Employee in both types I(Employee) and I(Works_On), and the same parameter Project in both types I(Works_On) and I(E→R). Thus, the parameters play a significant role by providing a means of representing linkages between two situation types, which, in an ER diagram, are linkages between different elements of its information content.

In addition, constructs in ER diagrams, such like 'Is-A' relationship, 'Ordinality' and 'Cardinality', domain, and attribute, can also be captured this way. Below we give another example taken from a live information system at one of our universities, an ER diagram as part of its conceptual design, in which we only show entities and relationships omitting attributes.

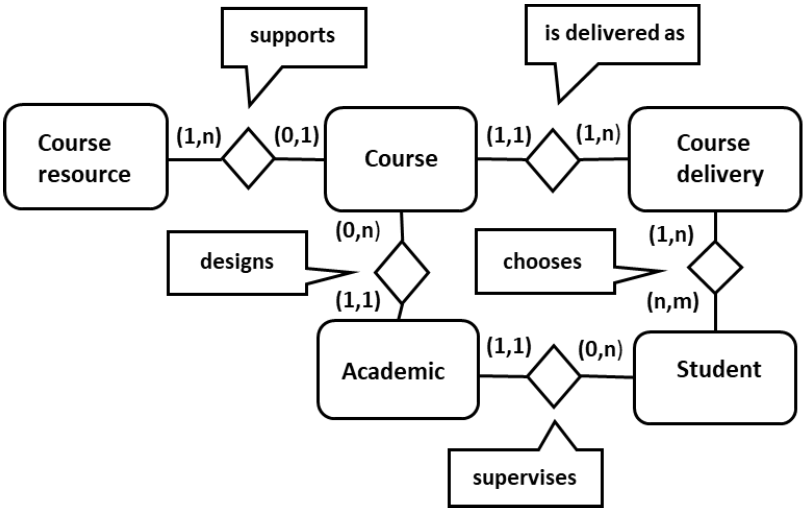

Figure 4 an ER Diagram as Part of a Live IS

The ER diagram in its simplified form (i.e., no attribute is included) shown in Figure 4 is part of the backend database of a system at the Beijing Union University in China that is concerned with course management. The real-world objects in question are academics, courses, resources for courses, the deliveries of a course, and students who choose and participate in the delivery of a course. The real-world facts are: 'an academic designs a course', 'a course is supported by various resources such as texts and software', 'a course is delivered more than once', 'a student chooses a course delivery', and 'an academic supervises a student'. The core of the information content of the ER diagram includes all of them, which is shown below:

I(Designs) = [s4 | s4 ⊨ <Designs, Academic, Course, L4, T4, 1>];
I(Supports) = [s5 | s5 ⊨ <Supports, Course_Resource, Course, L5, T5, 1>];
I(Is_delivered) = [s6 | s6 ⊨ < Is_delivered_as, Course, Course_Delivery, L6, T6, 1>];
I(Chooses) = [s7 | s7 ⊨ <Chooses, Student, Course_Delivery, L7, T7, 1>];
I(Supervises) = [s8 | s8 ⊨ <Designs, Academic, Course, L8, T8, 1>];

Note that all parameters for location and time are numbered differently in order to show that they are independent of one another.

6 Information Content Mapping

The definition of the information content of an ER diagram and the method to define it under the SBR framework that were described in the previous section help information content mapping. In this section, we will briefly introduce an approach to mapping information content between diverse ER diagrams as an application of our ideas presented in this paper. We use the construct called argument role (Devlin, 1991, p.116), which was discussed earlier in Definition 8 of Section 3, to generate mappings. Each n-place relation of an infon has a fixed collection of 'slots' into which appropriate objects can be placed, and these slots are called argument roles. Here is another example - we can specify the selling relation like this:

<sells | seller, buyer, object-sold, price, location, time>

in which seller, buyer etc. are argument roles.

Definition 11 (Information content mapping). let $\Phi_{(s)}$ and $\Gamma_{(T)}$ be the information contents of two ER diagrams *S* and *T* respectively, and let ω be a situation type in which argument roles are specified, for information units $\sigma \in \Phi_{(s)}$ and $\upsilon \in \Gamma_{(T)}$

$$\sigma \rightarrow \omega, \upsilon \rightarrow \omega \vDash \sigma = \text{IC-M}(\upsilon) \text{ and } \upsilon = \text{IC-M}(\sigma)$$

where '$\rightarrow$' means 'can be generalised to' and IC-M is an *information content mapping* function.

This definition says that if two situation types both can be generalised to one that has argument roles specified then the first two situation types map to each other.

Generation of a Common-Role Table

To facilitate the identification of the ‘‘can be generalised to’ relationships between situation types, we construct a so-called Common-Role table for the situation types that formulate the information content of diverse ER diagrams.

We use the following expression to denote a tuple in a Common-Role table:

<Role, Concept, DATABASE>.

Figure 5 shows two simple ER diagrams from two different data sources, Database A and Database B.

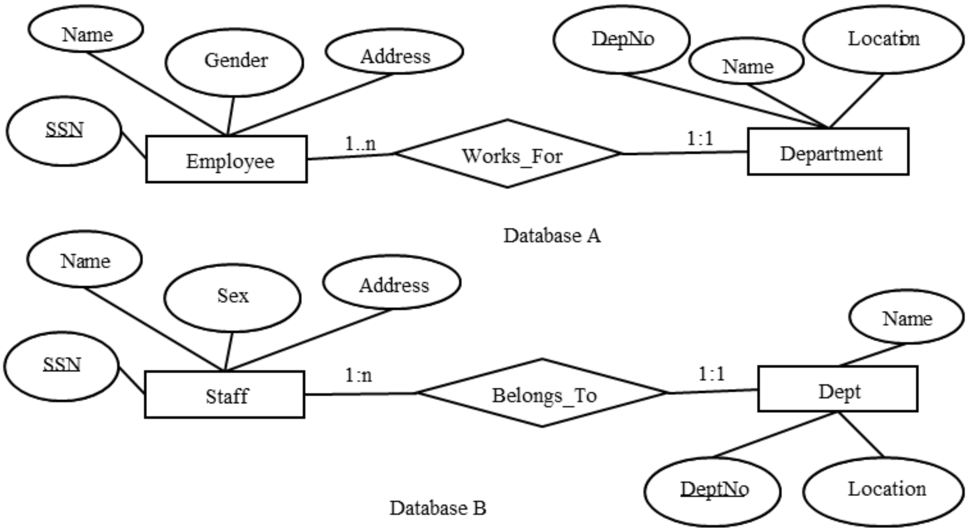

Figure 5 Example of two ER Diagrams

Then, we may have a Common-Role table displayed in Figure 6.

Role	**Concept**	**DATABASE**
Company_Person	Employee	Database A
Company_Person	Staff	Database B
Workplace	Department	Database A
Workplace	Dept	Database B
Employeed_By	Works_For	Database A
Employeed_By	Belongs_To	Database B
Company_Person_Gender _Gender	Gender	Database A
Company_Person_Gender _Gender	Sex	Database B
etc.		

Figure 6 A Common-Role Table

Generation of Information Content Mapping

Following Definition 11 and using the definition of argument role, mappings between the situation types in the information contents of different databases (in our case, the databases are represented by ER diagrams), can be generated.

Take Figure 5 as an example. A situation type, $\sigma \in \Phi_{(A)}$ where $\Phi_{(A)}$ is the information content of Database A, can be denoted like this:

$$\sigma = [s_1 \mid s_1 \models \langle \text{Works\_For, Employee, Department A}, 1\rangle].$$

According to the definition of argument role and ‘Role’ given in Figure 5, we re-write the above expression like this:

$$\sigma' = [s_1 \mid s_1 \models \langle \text{Employeed\_By:Works\_For, Company\_Person:Employee,} \\ \text{Workplace:Department, A}, 1\rangle].$$

In the same way, we shall have another situation type, $\upsilon \in \Gamma_{(B)}$ where $\Gamma_{(B)}$ is the information content of Database B, which can be denoted like this when argument roles are considered:

$$\upsilon' = [s_2 \mid s_2 \models \langle \text{Employeed\_By:Belongs\_To, Company\_Person:Staff,} \\ \text{Workplace:Dept, B}, 1\rangle].$$

Now, we can easily find out that both σ’ and υ’ can be generalised into the situation type ω:

$$\omega = \langle \text{REL}^2\text{:Employeed\_By, IND:Company\_Person, IND:Workplace}, 1\rangle.$$

Hereby, by Definition 11, we have

- $\sigma = \text{IC-M}(\upsilon)$ and
- $\upsilon = \text{IC-M}(\sigma)$

The above shows how relationships are mapped. In a similar way, entities can be mapped also. For example,

$$\sigma_1' = [s_{11} \mid s_{11} \models \langle \text{ Company\_Person\_Details:Employee\_Details,} \\ \text{Company\_Person\_Name:Employee\_name,}$$

Company_Person_SSN:Employee_SSN,Company_Person_Gender:
Employee_Gender, Company_Person_Address: Employee_Address, A, 1>].

υ_1' = [s_{21} | $s_{21} \models$ <Company_Person_Details:Staff_Details,
Company_Person_Name:Staff_name, Company_Person_SSN: Staff_SSN,
Company_Person_Gender: Staff_Sex, Company_Person_Address:
Staff_Address, B, 1>].

As we did for the relationships, we can generalise both σ_1' and υ_1' into the situation type ω_1:

ω_1 = <REL[4]: Company_Person_Details, IND: Company_Person_Name, IND:
Company_Person_SSN, IND: Company_Person_Gender, IND:
Company_Person_Address, 1>.

And again, by Definition 11, we have

- σ_1 = IC-M (υ_1) and
- υ_1 = IC-M (σ_1)

7 Conclusions and Future Work

In this paper, we have presented an approach to exploring and formulating the notion of 'information content' of an information system from a particular information-theoretic perspective. We draw on qualitative information theories including Dretske's semantic theory of information, Devlin's 'infon' theory and the information flow channel theory, and then put forward a framework consisting of information sources, bearers, and receivers. Under this framework, we described how the information content of an ER diagram may be formulated. We then applied these ideas to exploring relationships between different ER diagrams by showing how their information contents may be mapped to each other. Once the information content of every single ER diagram has been discovered and formulated, and information content mappings between ER diagrams have been generated, it can be exploited by the users for posing queries and building a knowledge base together with information-content inference rules, which we described in Feng and Hu, 2002. We can conclude that our work presented here is an attempt to find out whether the qualitative information theories that we follow are useful for exploring the notion of information content for information systems, and we find that they are enlightening and helpful.

We have shown that the SBR framework plays a pivotal role in our approach. We now summarise how the framework has helped and may further help understand information systems. First, the framework serves as an information-theoretic perspective to explore and then define the notion of information content of information systems, and particularly it enables us to identify and then formulate the information content of an ER diagram, which makes sense within the framework. Second, the framework can be seen as a particular systemic view for information systems, centred on the notions of information and information flow. Third, arguably this systemic view would help reveal the essence of information systems and that of the process of their development. Fourth and finally, the perspective and ideas embodied by the framework seem to be able to provide information system develop methodologies with some underpinning within which we have shown how it underpins database design.

We will continue with this line of the work by looking at other representations of data, such as object-oriented (O-O) schemata, relational schemata, and semi-structured data (e.g., XML). We will also explore other streams of information theories including Maturana and Varela's way of thinking about information (1992), which seems to put a far greater emphasis on the role that human beings play in approaching the concept of information. Our goal of this line of work is to develop an approach to adequately exploring and formulating the information content of an information system.

References

Andritsos, P., Miller, R. J. and Tsaparas, P. (2004) Information-Theoretic Tools for Mining Database Structure from Large Data Sets. SIGMOD Conference 2004.
Arenas, M. and Libkin L. (2003) An Information-Theoretic Approach to Normal Forms for Relational and XML Data PODS 2003, June 912, 2003, San Diego, CA.
Arens, Y., Chee, C.Y., Hsu, C. and Knoblock, C. A. (1993) "Retrieving and Integrating Data from Multiple Information sources" in International Journal of Intelligent and Cooperative Information Systems. Vol. 2, No. 2. Pg. 127-158, 1993.

Barwise, J. and Seligman, J. (1997) Information Flow – the logic of distributed systems. Cambridge University Press, Cambridge
Barwise, J. and Perry, J. (1983) Situations and Attitudes. Bradford Books, MIT Press.
Berlin, J. and Motro, A. (2002) Database Schema Matching Using Machine Learning with Feature Selection. In Proceedings of CAiSE-02, The 14th International Conference on Advanced Information Systems Engineering, Toronto, Canada, May 2002. Lecture Notes in Computer Science No. 2348, Springer-Verlag, Berlin, Germany, pp. 452--466.
Bernstein, P. A. and Haas, L. M. (2008) "Information Integration in the Enterprise", Communications of the ACM Vol. 51, No. 9, September 2008 https://doi.org/10.1145/1378727.1378745.
Blanco, J. M., Goni, A. and Illarramendi, A. (1999) Mapping Among Knowledge Bases and Data Repositories: Precise Definition of Its Syntax and Semantics. Information Systems, Vol.24, No. 4, pp. 275-301, 1999.
Bright, M.W., Hurson, A.R. and Pakzad, S. (1994) "Automated Resolution of Semantic Heterogeneity in Multidatabases". ACM Transactions on Database Systems, Vol.19, No.2, June 1994, pp.212-253.
Chawathe, S., Garcia Molina, H., Hammer, J., Ireland, K., Papakostantinou, Y., Ullman, J. and Widom, J. (1994) "The TSIMMIS project: Integration of Heterogeneous Information Sources" in IPSJ Conference, Tokyo, Japan, 1994.
Checkland, P. (1981) Systems Thinking, Systems Practice, John Wiley & Sons, Chichester.
Checkland, P. and Holwell, S. (1998) Information, Systems and Information System, John Wiley & Sons, Chichester, New York.
Devlin, K. (1991) Logic and Information. Cambridge University Press, Cambridge.
Dimopoulos, Y. and Kakas, A. (2001) Information Integration and Computational Logic. CoRR arXiv: cs. AI/ 0106025.
Dretske, F. I. (1981) Knowledge and the Flow of Information. Basil Blackwell, Oxford.
Feng, J. (1999) "The use of Semantic Theories of Information in Identifying and Formulating Information Requirements for Information Systems Design", in proceedings of BIT'99, Manchester.
Feng, J. and Hu, W. (2002) Inference rules for reasoning on informational content and algorithm for simplifying conceptual data schemas. Proceedings of the 7th Annual Conference of the UK Academy for Information Systems, UKAIS 2002,

Information Systems Research, Teaching and Practice, Published by Leeds Metropolitan University, Leeds, 2002. ISBN 1-898883-149, pp.158-166.
Floridi, L. (2002) "What Is the Philosophy of Information?" Metaphilosophy, 33(1–2): 123–145. doi:10.1111/1467-9973.00221.
Genesereth, M., Keller, A. and Dushka, O. (1997) Informaster: an information integration system. In Proceedings of the 1997 ACM SIGMOD.
Hammer, J. and Mcleod, D. (1993) "An Approach to Resolving Semantic Heterogeneity in a Federation of Autonomous, Heterogeneous Database Systems" International Journal of Intelligent and Cooperative Information Systems, 2:51-83, 1993.
Hull, R. (1984) "Relative Information Capacity of Simple Relational Database Schemata", Proceedings ACM Conference on Principles of Database Systems, 1984, 97-109.
Levene, M. (1998) "On the Information Content of Semi-Structured Databases", Acta Cybernetica, vol. 13, pp. 257-275.
Liu, K. (2000) Semiotics in information systems engineering. Cambridge University Press, Cambridge.
Langefors, B. (1973) Theoretical Analysis of Information Systems, Studentlitteratur, Auerbacher
Microsoft Power BI (2020). Microsoft Documentation.
https://docs.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview
Maturana, H. and Varela, F. (1992) The Tree of Knowledge, Shamb
Ala.
Miller, R. J., Ioannidis, Y. E. and Ramakrishnan, R. (1993) "The Use of Information Capacity in Schema Integration and Translation", in Proceedings of the 19th International Conference on Very Large Data Base, Morgan Kaufmann, San Francisco.
Mingers, J. (1995) Information and meaning: foundations for an intersubjective account, Information Systems Journal, Volume5, Issue4 October 1995, Pages 285-306 First published: October 1995 https://doi.org/10.1111/j.1365-2575.1995.tb00100.x.
Rosenthal, A. and Reiner, D. S. (1994) Tools and Transformations - Rigorous and Otherwise - for Practical Database Design. ACM Trans. Database Syst. 19(2): 167-211 (1994).
Shannon, C. E. and Weaver, W. (1949) The mathematical theory of communication. Urbana: University of Illinois.

Stamper, R. (1997) "Organisational Semiotics", In Information Systems: An Emerging Discipline? Mingers, J and Stowell, F. ed. The McGraw-Hill Companies, London.
Stanford Encyclopedia of Philosophy (2018) entry Information First published Fri Oct 26, 2012; substantive revision Fri Dec 14, 2018.
Stowell, F. and Welch, C. (2012) The Manager's Guide to Systems Practice: Making Sense of Complex Problems. Hohn Wiley & Sons ltd, United Kingdom.
Stowell, F. and West, D. (1994) Client-led Design, McGraw-Hill, London.
Tableau Software (2020) https://www.tableau.com/en-gb.
Weick, K. (1995) Sensemaking in Organizations. SAGE Publications, Inc, ISBN 13: 9780803971776.

UK SYSTEMS SOCIETY

UKSS Gold Medallists

Professor Russell Ackoff
Professor Stafford Beer
Dr Fritjof Capra
Professor Peter Checkland
Professor C. West Churchman
Professor Humberto Maturana
Sir Geoffrey Vickers

***Systemist* is a publication of**

The United Kingdom Systems Society