CHAPTER 2

DATA AND INFORMATION

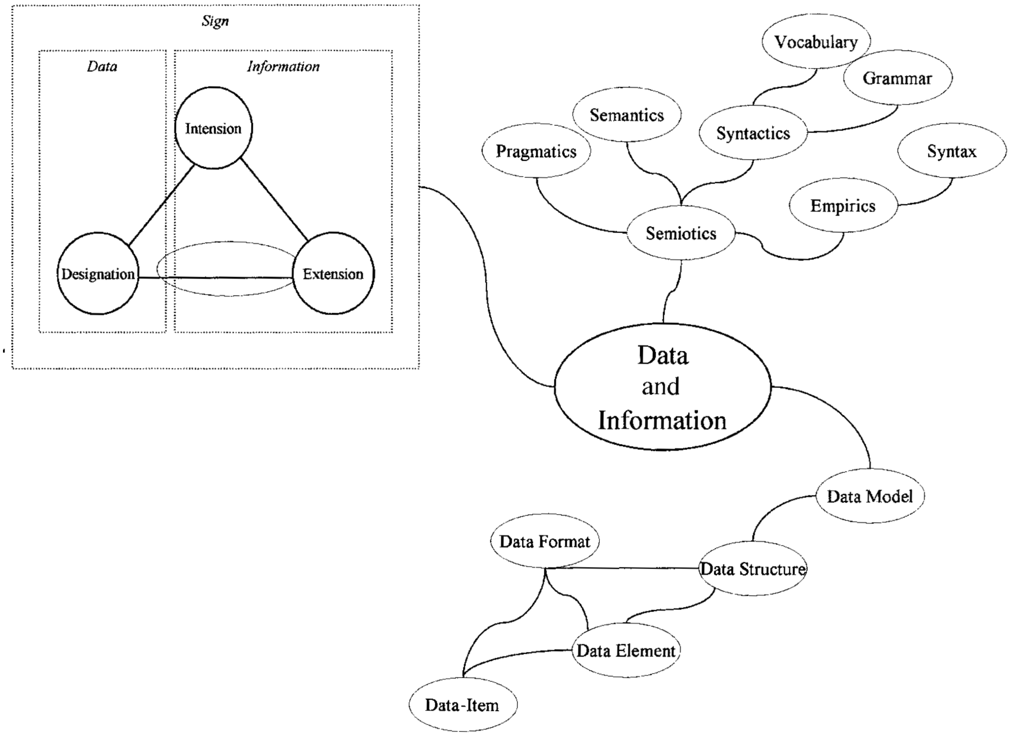
It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.

Sir Arthur Conan Doyle (1859–1930)

LEARNING OUTCOMES

At the end of this chapter the reader will be able to:

- Distinguish between data and information in terms of the concept of a sign
- Describe the four layers of a sign
- Distinguish between syntax and semantics, and explain the relevance of this distinction to the concept of a database schema



2.1 INTRODUCTION

In Chapter 1 we defined data as facts or positive assertions about some domain of discourse. According to this definition a database is a logically organised collection of facts about some domain. A datum – a unit of data – is a symbol or a set of symbols which is used to represent something. This relationship between symbols and what they represent is the essence of what we mean by information. Hence, information is interpreted data – data supplied with semantics. In this chapter we use the concept of a sign and the discipline which studies signs – semiotics – to help us more clearly explain the distinction between data and information.

Much of database work, and indeed information systems work, takes place without an examination of this critical distinction between data and information. Hence, readers may skip this chapter without fear of being unable to develop effective database systems. However, an understanding of this distinction is important for a number of reasons:

- It has a bearing on the nature of database development. We shall argue that requirements specification techniques such as entity–relationship diagramming (Chapter 16) are sign-systems
- It helps us understand more clearly the place of database systems within ICT systems and information systems more generally (Chapter 4)
- It helps us understand the critical place of data in modern business and the need for effective planning (Chapter 21) and administration (Chapter 23) of data.

2.2 SIGNS

The concept of information is an extremely vague one, open to many different interpretations (Stamper, 1989). One conception popular in the computing literature is that information results from the processing of data: the assembly, analysis or summarisation of data. This conception of information as analogous to chemical distillation is useful, but ignores the important place of human interpretation in any understanding of information.

We shall argue that both data and information are embodied in the concept of a sign and that hence signs and sign-systems are the fundamental stuff of database systems.

A sign is anything that is significant. In a sense, everything that humans do is significant to some degree. The world within which humans find themselves is resonant with sign-systems. A sign-system is any organised collection of signs. Everyday spoken language is probably the most readily accepted and complex example of a sign-system. Signs however exist in most other forms of human activity, since signs are critical to the process of human communication and understanding.

Example Humans communicate through non-verbal as well as verbal sign-systems. We colloquially refer to such non-verbal communication as 'body language'. Hence, humans can impart a great deal in the way of information by facial movements and other forms of bodily gesture. Such gestures are also signs.

Note the link between the words *sign* and *significant* in English. These words clearly have the same root. The concept of the significance of signs cannot be divorced from people. Different people find different things significant. Many such differences in interpretation are due to differences in the context and culture of communication.

2.3 SEMIOTICS

Semiotics or semiology is the study of signs or more precisely of sign-systems. Signs and sign-systems can be considered in terms of four interdependent levels (Stamper, 1989): pragmatics, semantics, syntactics and empirics. These constitute the four main branches of semiotics.

2.3.1 PRAGMATICS

Pragmatics is the study of the general *context and culture* of communication or the shared assumptions underlying human understanding. For communication to occur between human beings signs must exist in a context of shared understanding. As we shall see, there must be agreed expectations among a group of people about the symbols and the referents or concepts they signify. Pragmatics is the study of such mutual understanding. Much of pragmatics can be considered as the study of culture – the common expectations underlying human communicative behaviour in a particular context.

2.3.2 SEMANTICS

Semantics is the study of the *meaning* of signs – the association between signs and behaviour. Semantics can be considered as the study of the link between symbols and their referents or concepts, particularly the way in which signs relate to human behaviour embodied in norms.

2.3.3 SYNTACTICS

Syntactics is the study of the logic and *grammar* of sign systems. Syntactics is devoted to the study of the physical form rather than the content of signs.

2.3.4 EMPIRICS

Empirics is the study of the *physical characteristics* of the medium of communication. Empirics is devoted to the study of communication channels and their characteristics, e.g. sound, light, electronic transmission etc.

2.4 SEMANTICS

The relationship between information and data is located in the area of meaning – semantics. Semantics is the study of what signs refer to. Communication involves the use and interpretation of signs. When we communicate, the sender has to externalise his or her intentions in terms of some signs. In face-to-face conversation this will involve the use of linguistic signs. The receiver of the message must interpret the signs. In other words, he or she must assign some meaning to the signs of the message. Semantics is concerned with this process of assigning meaning to signs.

A simple model of semantics is one in which a sign can be broken down into three component parts, which are frequently referred to collectively as the meaning triangle (Sowa, 1984) (see Figure 2.1):

- *Designation.* The designation of a sign refers to the symbol (or collection of symbols) by which some concept is known. The designation of a sign is sometimes referred to as the signifier – that which is signifying something

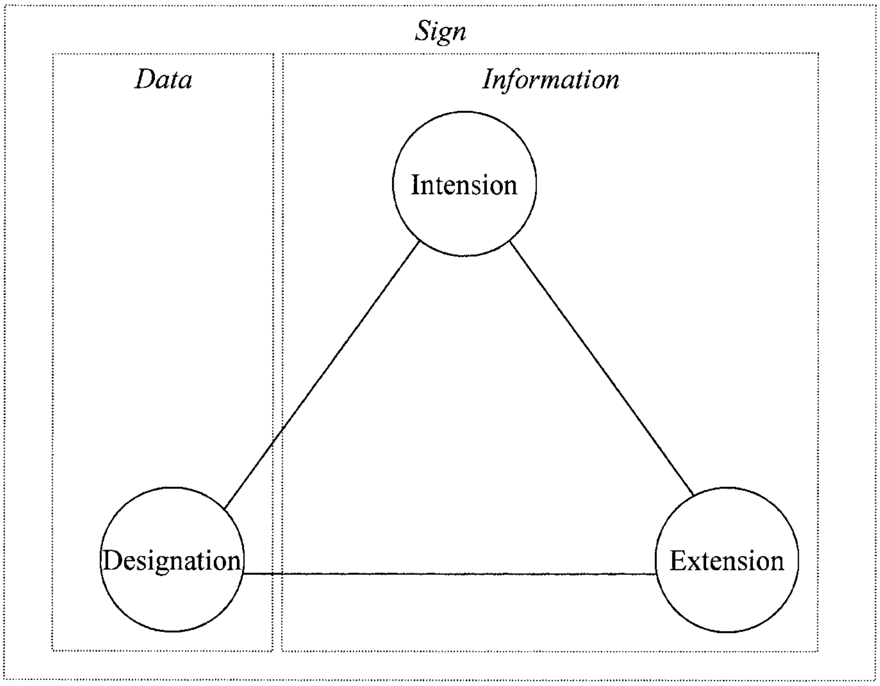


Figure 2.1 The meaning triangle.

- *Extension*. The extension of a sign refers to the range of phenomena that the concept in some way cover. The extension is sometimes known as the referent or the signified – that which is being signified
- *Intension*. The intension of a sign is the collection of properties that in some ways characterise the phenomena in the extension. The intension is the idea of significance

The designation or symbols are equivalent to data in the classic language of information systems. A datum, a single item of data, is a set of symbols used to represent something. Information particularly occurs in the 'stands-for' relation between the designation and its intension, and the intension and its extension.

Examples ➡ In a manufacturing database system the symbols 43 constitute the designation. A possible extension is a collection of products. The intension may be the quantity of a product sold.

The symbols M and F might be significant in some informatics context. To speak of information we must supply some intension and/or extension for the symbols. M might have as its extension the male population; F might have as its extension the female population. Taken together, the meaning of these symbols is supplied by the intension of human gender.

The meaning triangle should not be taken to mean that signs have an inherent meaning. A sign can mean whatever a particular social group chooses it to mean. The same sign may mean different things in different social contexts. As interpreters of signs humans are extremely proficient at assigning the correct interpretation for a sign in a particular context.

Examples ➡ In the Welsh language the same verb *dysgu* (pronounced dusgey) is used both for *to teach* and *to learn*. Hence the same sentence – *Rydw i'n dysgu* – can mean either *I am learning* or *I am teaching* depending on the context supplied usually by elements such as the rest of the conversation.

Take the sign 030500. In a given context we interpret this as a date. However, in the UK we would interpret the significance of the sequence of digits differently from those in the USA. In the UK the first two digits represent the day of the month. In the USA the first two digits represent the month.

2.5 SYNTACTICS

The most important category of sign-system used in human communication is that of language. Syntactics concerns itself with all these representational aspects of language. All languages consist of the following elements:

- *Vocabulary*. A complete list of the terms of a language
- *Grammar*. Rules that control the correct use of a language
- *Syntax*. The operational rules for the correct representation of terms and their use in the construction of sentences of the language

We may distinguish between two broad categories of languages: natural languages and formal languages:

- *Natural languages*. These are languages used for everyday communication and include languages such as English, French and Welsh. Natural languages have evolved over time in particular linguistic communities, and indeed continue to evolve. As such they tend to be extremely rich and complex in terms of vocabulary, grammar and syntax
- *Formal languages*. These are artificial languages normally constructed for some defined purpose. As such they tend to have much simpler vocabularies, grammar and syntax. Work in the information systems area abounds with a variety of such formal languages

Example ➡ All the existing programming languages such as C++ or Java are examples of formal languages.

Some of the most well-developed examples of formal languages are those developed in formal logic, including propositional and predicate logic. These languages are useful as ways of representing restricted universes of discourses and developing reasoning about their properties.

Example ➡ Many of the formal languages used to build database systems, such as structured query language (SQL) (Part 3), are founded in formal logic.

2.6 SCHEMAS

The term universe of discourse (UoD), or domain of discourse, is sometimes used to describe a context within which a group of signs (usually linguistic terms) is used continually by a social group or groups. For work in the area of information systems it is important to develop a detailed understanding of the structure of these terms. In database work this structure is known as a schema. A schema is an attempt to develop an abstract description of some UoD, usually in terms of a formal language such as logic.

In classic Aristotelian logic the intension of a sign is a collection of properties that may be divided into two groups: the defining properties that all phenomena in the extension must have; and the properties that the phenomena may or may not have. The nature of the defining properties is such that it is objectively determinable whether or not a property has certain phenomena. In classical logic, such properties are designated by predicates (Chapter 9). Hence, defining properties in this view determines a sharp boundary between membership and non-membership of a sign.

A database schema is made up of a series of signs. A schema is therefore a representation of some reality. Although there are other interpretations of the link between intension and extension, in traditional terms, signs in a schema adhere to an Aristotelian interpretation. In a database we must be able to provide an unambiguous procedure for determining whether a phenomenon is a member of a sign or not.

2.7 DATA STRUCTURES, DATA-ITEMS, DATA ELEMENTS AND DATA FORMATS

To build a schema we must have a system of signs or a language we can use. Such a language must have an agreed vocabulary, grammar and syntax. The formal language used for defining schemas is generally described as a data model. A data model establishes a set of principles for organising data. This set of principles that define a data model may be divided into three major parts:

- *Data definition.* A set of principles concerned with how data is structured
- *Data manipulation.* A set of principles concerned with how data is operated upon
- *Data integrity.* A set of principles concerned with determining which states are valid for a database

In general terms, the syntax of data definition in any data model tends to use a hierarchy of data-items, data elements and data structures (Tsitchizris and Lochovsky, 1982).

- A data-item is the lowest-level of data organisation. A data-item is the atomic construct in any data model. In other words, that construct that cannot be divided any further
- A data element is a logical collection of data-items
- A data structure is a logical collection of data elements

Example Hence, in the relational data model (Chapter 7), a data-item corresponds to a column of a table. A data element corresponds to a row of a table, and the one and only data structure in the relational data model is the table or relation.

Some of the semantics of data-items, data elements and data structures in a data model is provided by the concept of a data format. The format of a data-item tends to be referred to as a data type. Some of the meaning of a given data-item is embodied in its data type. A data type declares the range of valid operations that is possible on a data-item.

Example We might declare a given data-item to be a person's age and this data-item to be declared on an integer data type. This means it is possible to conduct the range of arithmetic operations on a person's age.

Most data needed in commercial information systems is what we might call standard data. Standard data is defined in terms of a number of standard data types. A data type acts as a categorisation of data and defines not only the format of some data-item but also the allowable set of operations we may perform on the data in the item. There are a number of standard data types used by most information systems applications. These include:

- *Text*. Strings of symbols made up of characters from the alphabet and a range of other characters
- *Numbers*, including integer, decimal and real numbers
- *Units of time*, including dates, seconds, minutes and hours

Example Clearly it is important for a computer system to know the type of data it is dealing with. For instance, it makes sense to add two integers such as 1 and 2 together. It makes very little sense to add the character A to the character B. Hence, a number data type will allow addition but a text data type will disallow it.

Information systems are now being used to capture, store and manipulate far more complex data types than the standard data types discussed above. This is to enable such systems to handle different media. Such complex data types include:

- *Images*. Graphics and photographic images
- *Audio*. Various forms of sound data
- *Video*. Various forms of moving image

Example These forms of complex or multi-media data need different coding schemes or formats to enable the representation of this data in digital form. For instance, images are coded in terms of pixels. A pixel is a picture element. A high-resolution computer monitor can be considered as a 1024 by 768 grid in which each cell of the grid is a pixel. To represent an image we therefore need to store, for each pixel, its colour in the image. For a complete monitor over 700,000 pixels are needed.

Information systems are now also required to handle complex data structures as well as complex data-items or elements. Organisations want to be able to define the structure of documentation that they use for communication. A document such as an invoice or a contract is a complex data structure in that it may be made up of text, numeric data and images. Also different organisations may use different formats for such documents. Hence, there is pressure on organisations in the same industrial sector to develop standards for such documentation. Such standards are being specified using a formal language known as XML – extensible markup language. XML is a variant of the Internet language HTML and is discussed in more detail in Chapter 39.

2.8 CASE STUDY: UNIVERSITY MODULE DESCRIPTIONS

Information is important to a university for successful operation and will include information about students, staff and modules. A module description is a critical sign-system in this environment. As an organisational communication a module description can be analysed on a number of levels:

- *Pragmatics* – the study of the general *context and culture* of communication or the shared assumptions underlying human understanding. Key assumptions surrounding the use of module descriptions are that the educational experience can be packaged or chunked in terms of modules, and delivered and assessed in discrete units. Effectively modules become the building blocks of the higher education experience. They are the educational products delivered by universities

- *Semantics* – the study of the *meaning* of signs or sign-systems. Within a university setting, students and staff tend to treat module descriptions as informal contracts. They define particularly what is to be covered in terms of the delivery of the educational experience. They also define the likely mechanism to be used for assessment of students

- *Syntactics* – the study of the logic and *grammar* of sign-systems. At the syntactic level a module description can be considered as a data structure. A standard format for module descriptions will normally be formulated at a particular university and will include data elements such as a module code, module name, level, learning outcomes, syllabus, assessment pattern and suggested reading. Data-items within such elements are likely to store characters of written text as key symbols

- *Empirics* – the study of the *physical characteristics* of the medium of communication. A module description may be made available in a physical form on paper as part of a student or module handbook. In the modern university it may also be made available electronically over a university intranet and may be stored within a database system for easy retrieval and maintenance

2.9 SUMMARY

- A database can be considered as an organised collection of data which is meant to represent some universe of discourse
- Data are facts. A datum, a unit of data, is one symbol or a collection of symbols that is used to represent something
- Data by itself is meaningless. To prove useful data must be interpreted. Information is interpreted data. Information is data placed within a meaningful context. Information is data with an assigned semantics or meaning
- The distinction between data and information is embodied in the concept of a sign
- Semiotics or semiology is the study of signs, or more precisely of sign-systems
- Signs and sign-systems can be considered in terms of four interdependent levels: pragmatics, semantics, syntactics and empirics
- The term universe of discourse (UoD), or domain of discourse, is sometimes used to describe a context within which a group of signs (usually linguistic terms) is used continually by a social group or groups. For work in the area of information systems it is important to develop a detailed understanding of the structure of these terms. In database work this structure is known as a schema. A schema is an attempt to develop an abstract description of some UoD, usually in terms of a formal language
- To build a schema we must have a system of signs or a language we can use. Such a language must have an agreed vocabulary, grammar and syntax. The formal language used for defining schemas is generally described as a data model
- In general terms, the syntax of data definition in any data model tends to use a hierarchy of data-items, data elements and data structures

2.10 ACTIVITIES

1. Analyse some organisational communication known to you in terms of the levels of semiotics. For instance, in what ways can a student grade for an assessment be treated as a sign? What does it signify in terms of pragmatics, semantics, syntactics and empirics?

2. After reading Part 3, attempt to detail some of the vocabulary, grammar and syntax of the formal language SQL. For instance, what is the syntax of the SELECT statement?
3. Use the meaning triangle to analyse an item of data – order number, debit from a bank account – in terms of the concept of a sign. For instance, what is the designation, intension and extension of an order number?

2.11 REFERENCES

Sowa, J.F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA, Addison-Wesley.

Stamper, R.K. (1989). *Information in Business and Administrative Systems*. London, Batsford.

Tsitchizris, D.C. and F.H. Lochovsky (1982). *Data Models*. Englewood Cliffs, NJ, Prentice-Hall.