Data

The word *data* (sing. datum) is originally Latin for “things given or granted”. Because of such a humble and generic meaning, the term enjoys considerable latitude both in its technical and in its common usage, for almost anything can be referred to as a “thing given or granted” (Cherry [1978]). With some reasonable approximation, four principal interpretations may be identified in the literature. The first three captures part of the nature of the concept and are discussed in the next section. The fourth is the most fundamental and satisfactory, so it is discussed separately, in section three. On its basis, some further clarifications about the nature of data are introduced in sections four to six. A reminder about the social, legal and ethical issues raised by the use of data concludes this entry.

Three interpretations of the concept of data

According to the *epistemic* (i.e. knowledge-oriented) interpretation, data are collections of *facts*. In this sense, data provide the basis for further reasoning – as when one speaks of data as the *givens* of a mathematical problem – or represent the basic *assumptions* or empirical *evidence* on which further evaluations can be based, as in a legal context. The limits of this interpretation are mainly two. First, it is too restrictive, as it fails to explain, for example, processes such as *data compression* (any encoding of data that reduces the number of data units to represent some unencoded data; see Sayood [2006] for an introduction) or *data cryptography* (any procedure used to transform available data into data accessible only by their intended recipient; see Singh [1999] for an introduction), which may apply to facts only in a loosely metaphorical sense. Second, it trades one difficult concept (data) for an equally difficult one (facts), when actually facts are more easily understood as the outcome of some data processing. For example, census data may establish a number of facts about the composition of a population.

According to the *informational* interpretation, data are *information*. In this sense, for example, *personal data* are equivalent to information about the corresponding individual. This interpretation is useful in order to make sense of expressions such as *data mining* (information gathering; see Han and Kamber [2001] for an introduction) or *data warehouse* (information repository). However, two major shortcomings show its partial inadequacy. First, although it is important to stress how information depends on data, it is common to understand the former in terms of the latter, not vice versa: information is meaningful and truthful data, e.g. “paper is inflammable” (Floridi [2003]). So one is left with the problem of understanding what data are in themselves. Second, not all data are informational in the ordinary sense in which information is equivalent to some content (e.g. a railway timetable) about a referent (e.g. the schedule of trains from Oxford to London). A music CD may contain gigabytes of data, but no information about anything (Floridi [2005]).

According to the *computational* interpretation, data are collections (sets, strings, classes, clusters etc.) of *binary elements* (digits, symbols, electrical signals, magnetic patterns, etc.) processed and transmitted electronically by technologies such as computers and cellular phones. This interpretation has several advantages. It explains why pictures, music files or videos are also constituted by data. It is profitably related both to the informational and to the epistemic interpretation, since a binary format is increasingly often the only one in which experimental observations or raw facts may be available and further manipulated (collected, stored, processed etc.) to generate information, e.g. in the course of scientific investigations (Baeyer [2003]). Finally, it highlights the malleable nature of data and hence the possibility of their automatic processing (Pierce [1980]). The main limit of this interpretation lies in the confusion between data and the format in which data may be encoded. Data need not be discrete (digital) they can also be analogue (continuous). A CD and a vinyl both contain music data. Binary digits are only the most recent and common incarnation of data.

Given the previous interpretations, it seems wise to exercise some flexibility and tolerance when using the concept of data in different contexts. On the other hand, it is interesting to note that the aforementioned interpretations all presuppose a more fundamental definition of data, to which we can not turn.

The diaphoric interpretation of data

A good way to uncover the most fundamental nature of data is by trying to understand what it means to erase, damage or loose them. Imagine the page of a book encrypted or written in a language unknown to us. We have all the data, but we do not know their meaning, hence we have no information or facts or evidence, yet. Suppose the data are continuous pictograms. We still have all the data, but no binary bits. Let us now erase half of the pictograms. We may say that we have halved the data as well. If we continue in this process, when we are left with only one pictogram we might be tempted to say that data require, or may be identical with, some sort of representations. But now let us erase that last pictogram too. We are left with a white page, yet not without data. For the presence of a white page is still a datum, as long as there is a difference between the white page and the page on which something is written. Compare this to the common phenomenon of "silent assent": silence, or the lack of perceivable data, is as much a datum as the presence of some rumour, exactly like the zeros of a binary system. We shall return to this point presently, but at the moment it is sufficient to grasp that a genuine, complete erasure of all data can be achieved only by the elimination of all possible differences. This clarifies why a datum is ultimately reducible to just a lack of uniformity. More formally, according to the *diaphoric interpretation* (*diaphora* is the Greek word for "difference"), the general definition of a datum is:

D) datum = x being distinct from y;

where the x and the y are two uninterpreted variables and the domain is left open to further interpretation.

This definition can be applied at three levels.

1) data as diaphora *de re*, that is, as lacks of uniformity in the world (Seife [2006]). There is no specific name for such "data in the wild". A possible suggestion is to refer to them as *dedomena* ("data" in Greek; note that the word "data" comes from the Latin translation of a work by Euclid entitled *Dedomena*). Dedomena are not to be confused with environmental data. They are pure data or proto-epistemic data, that is, data before they are interpreted. They can be posited as an external anchor of information, for dedomena are never accessed or elaborated independently of a level of abstraction. They can be reconstructed as requirements for any further analysis: they are not experienced but their presence is empirically inferred from (and required by) experience. Of course, no example can be provided, but data as dedomena are whatever

lack of uniformity in the world is the source of (what looks to information systems like us as) data, e.g. a red light against a dark background.

2) data as diaphora *de signo*, that is, lacks of uniformity between (the perception of) at least two physical states of a system, such as a higher or lower charge in a battery, a variable electrical signal in a telephone conversation, or the dot and the line in the Morse alphabet.

3) data as diaphora *de dicto*, that is, lacks of uniformity between two symbols of a code, for example the letters A and B in the Latin alphabet.

Depending on one's interpretation, dedomena in (1) may be either identical with, or what makes possible signals in (2), and signals in (2) are what make possible the coding of symbols in (3).

The dependence of information on the occurrence of well-structured data, and of data on the occurrence of differences (dedomena) variously implementable physically, explain why information can so easily be decoupled from its support. The actual *format*, *medium* and *language* in which data (and hence information) are encoded is often irrelevant and hence disregardable. In particular, the same data may be analog or digital, printed on paper or viewed on a screen, in English or in some other language, expressed in words or pictures, quantitative or qualitative.

Interpretations of the support-independence of data can vary quite radically, for the definition (D) above leaves underdetermined

- the classification of the relata (*taxonomic neutrality*);
- the logical type to which the relata belong (*typological neutrality*); and
- the dependence of their semantics on a producer (*genetic neutrality*).

We shall now look at each form of neutrality in turn.

Taxonomic neutrality

A datum is usually classified as the entity exhibiting the anomaly, often because the latter is perceptually more conspicuous or less redundant than the background conditions. However, the relation of inequality is binary and symmetric. A white sheet of paper is not just the necessary background condition for the occurrence of a black dot as a datum, it is a constitutive part of the [black-dot-on-white-sheet] datum itself, together with the fundamental relation of inequality that couples it with the dot. Nothing is a datum in itself. Rather, being a datum is an external property. This is summarised by the principle of taxonomic neutrality:

TaN) a datum is a relational entity.

The slogan is "data are relata", but the definition of data as differences is neutral with respect to the identification of data with *specific* relata. In our example, one may refrain from identifying either the red light or the white background as the datum.

Typological neutrality

Five classifications of different types of data as relata are quite common. They are not mutually exclusive and, depending on circumstances, on the sort of analysis conducted and on the level of abstraction adopted, the same data may fit different classifications.

1) *Primary data*. These are the principal data stored e.g. in a database, for example a simple array of numbers. They are the data an information-management system is generally designed to convey (in the form of information) to the end user. Normally,

when speaking of data one implicitly assumes that *primary* data is what is in question. So, by default, the red light of the low battery indicator flashing is assumed to be an instance of primary data conveying primary information.

2) *Secondary data*. These are the converse of primary data, constituted by their absence. Clearly, silence may be very informative. This is a peculiarity of data: their absence may also be informative.

3) *Metadata*. These are indications about the nature of some other (usually primary) data. They describe properties such as location, format, updating, availability, usage restrictions, and so forth. Correspondingly, *metainformation* is information about the nature of information. "'Rome is the capital of Italy' is encoded in English" is a simple example.

4) *Operational data*. These are data regarding the operations of the whole data system and the system's performance. Correspondingly, *operational information* is information about the dynamics of an information system. Suppose a car has a yellow light that, when flashing, indicates that the car checking system is malfunctioning. The fact that the light is on may indicate that the low battery indicator is not working properly, thus undermining the hypothesis that the battery is flat.

5) *Derivative data*. These are data that can be extracted from some data whenever the latter are used as indirect sources in search of patterns, clues or inferential evidence about other things than those directly addressed by the data themselves, e.g. for comparative and quantitative analyses. For example, from someone's credit card data, concerning e.g. the purchase of petrol in a certain petrol station, one may infer the derivative information of her whereabouts at a given time.

Let us now return to our question: can there be dataless information? The definition of data given above in (D) does not specify which types of relata are in question, only that data are a matter of a relation of difference. This *typological neutrality* is justified by the fact that, when the apparent absence of data is not reducible to the occurrence of *negative* primary data, what becomes available and qualifies as information is some further non-primary information x about y constituted by some non-primary data z. For example, if a database query provides an answer, it will provide at least a *negative* answer, e.g. "no documents found". This datum conveys primary negative information. However, if the database provides no answer, either it fails to provide any data at all, in which case no specific information is available – so the rule "no information without data" still applies – or it can provide some data to establish, for example, that it is running in a loop. Likewise, silence, this time as a reply to a question, could represent negative primary information, e.g. as implicit assent or denial, or it could carry some non-primary information, e.g. about the fact that the person has not heard the question, or about the amount of noise in the room.

Genetic neutrality

Finally, let us consider the semantic nature of the data. How data can come to have an assigned meaning and function in a semiotic system in the first place is one of the hardest problems in semantics. Luckily, the point in question here is not *how* but *whether* data constituting information as semantic content can be meaningful *independently* of an informee. The *genetic neutrality* (GeN) principle states that:

GeN) data can have a semantics *independently* of any informee.

Before the discovery of the Rosetta Stone, Egyptian hieroglyphics were already regarded as information, even if their semantics was beyond the comprehension of any interpreter. The discovery of an interface between Greek and Egyptian did not affect the semantics of the hieroglyphics but only its accessibility. This is the weak sense in which meaningful data may be embedded in information-carriers informee-independently. GeN supports the possibility of *information without an informed subject* and it is to be distinguished from the stronger, realist thesis (supported for example by Dretske [1981]), according to which data could also have their own semantics independently of an intelligent *producer/informer*.

Conclusion

Large part of social research involves the study of logical relationships between sets of attributes (variables). Some of these variables are dependent. They represent the facts that a theory seeks to explain. Some other variables are independent. They are the data on which the theory is developed. Thus, data are treated as factual elements that provide the foundation for any further theorising. It follows that data observation and collection and data analysis are fundamental processes to elaborate a theory, and computational social science (high-performance computing, very large data storage systems, and dedicate software for fast and efficient data collection and analysis) has become an indispensable tool for the social scientist. This poses several challenges. Some are technical. For example, data may result from a variety of disparate sources (especially when collected through the Internet), whose reliability needs to be checked; may be obtainable only through sophisticated processes of data mining and analysis whose accurate functioning needs to be under constant control; or the scale and complexity and heterogeneous nature of the dataset may pose daunting difficulties, computationally, conceptually and financially. Some other challenges are intellectual, ethical, political or indeed social in themselves. In this case, quality control (e.g. timely, updated and reliable data), availability (e.g. which and whose data are archived, and through what tools), accessibility (e.g. old codification systems or expensive fees can make available data practically inaccessible; privacy issues need to be taken into serious consideration), centralization (e.g. economy of scale, potential synergies, the increased value of large databases), political control (e.g. who exercises what power over which available data sets and their dissemination) are only some of the main issues that determine the initial possibility and final value of social research.

Data are the sap of any information system and any social research that relies on it. Their corruption, wantonly destruction, unjustified concealment, illegal or unethical use may easily undermine the basic processes on which not only scientific research but also the life of individuals and their complex societies depends (Brown and Duguid [2002]). In light of their importance, their whole life cycle – from collection or generation through storage and manipulation to usage and possible erasure – is often protected, at different stages, by legal systems in various ways and in many different contexts. Examples include copyright and ownership legislation, patent systems, privacy protection laws, fair use agreements, regulations about availability and accessibility of sensitive data, and so forth. The more societies develop into data-based societies, the more concerned and careful they need to become about their very foundation.

Unsurprisingly, in recent years a new area of applied ethics, known as information ethics (Floridi [1999]), has begun to address the challenging ethical issues raised by the new data-based environment in which advanced societies grow.

Luciano Floridi

Fellow of St Cross College, University of Oxford

Professor of Logic and Epistemology, Università degli Studi di Bari, Italy

Word count 2848 (references below and this note excluded)

References

Baeyer, H. C. v. 2003, *Information : The New Language of Science* (London: Weidenfeld & Nicolson).

Brown, J. S., and Duguid, P. 2002, *The Social Life of Information* (Boston: Harvard Business School Press). Paperback edition of the 2000 hardback edition, includes some corrections and revisions.

Cherry, C. 1978, *On Human Communication : A Review, a Survey, and a Criticism* 3rd ed (Cambridge, Mass ; London: MIT Press).

Dretske, F. I. 1981, *Knowledge and the Flow of Information* (Oxford: Blackwell). Reprinted in 1999 (Stanford, CA: CSLI Publications).

Floridi, L. 1999, "Information Ethics: On the Philosophical Foundations of Computer Ethics", *Ethics and Information Technology*, 1(1), 37-56.

Floridi, L. 2003, "Information" in *The Blackwell Guide to the Philosophy of Computing and Information*, edition, edited by L. Floridi (Oxford - New York: Blackwell), 40-61.

Floridi, L. 2005, "Is Information Meaningful Data?" *Philosophy and Phenomenological Research*, 70(2), 351-370.

Han, J., and Kamber, M. 2001, *Data Mining : Concepts and Techniques* (San Francisco, Calif. ; London: Morgan Kaufmann).

Pierce, J. R. 1980, *An Introduction to Information Theory : Symbols, Signals & Noise* 2nd edition (New York: Dover Publications).

Sayood, K. 2006, *Introduction to Data Compression* 3rd (Amsterdam ; London: Elsevier).

Seife, C. 2006, *Decoding the Universe: How the New Science of Information Is Explaining Everything in the Cosmos, from Our Brains to Black Holes* (New York: Viking).

Singh, S. 1999, *The Code Book : The Science of Secrecy from Ancient Egypt to Quantum Cryptography* (London: Fourth Estate).