Chapter 2
What is Information?

Abstract Chapter 2 examines how the most current use of the word 'information' can lead to outline an axiomatic definition of information. Its most specific features are that it has no existence unless it is physically inscribed as a sequence of symbols on some medium, which however has no influence on it besides ensuring its existence, and that it can be defined only as an equivalence class, with respect to transformations like alphabet change and coding. It is thus an abstract entity which resides in the physical world. An information meets Barbieri's concept of 'nominable entity', which refers to a singular object. This concept is explicated and illustrated. A natural number can be used, besides its usual meanings of representing a quantity (cardinal number) or a rank in a sequence (ordinal number), as a label uniquely representing a nominable entity. The uniqueness of nominable entities entails that their representatives do not suffer any change and thus must be protected against any perturbation. A short history of communication engineering, which developed the means of such a protection referred to as 'error-correcting codes', is briefly presented. It is also stated that the theoretical tools needed in order to deal with communication at a distance can be used as well for communication over time such as biological heredity.

2.1 Information in a Usual Meaning

We think it is helpful to begin a discussion of the information concept with examining a usual meaning of the word 'information' so as to determine what of its features endow it with the status of a scientific entity. Maybe the most familiar modern use of the word concerns mass media where an information consists of telling that some event has occurred and/or of reporting its circumstances. Some source then transmits some spoken or written text, or a succession of sounds and/or images, i.e., some *message*, in order to let know something to some recipient. A characteristic feature of an information is that it is *new* for its recipient or, more precisely, that a message is perceived as an information only if it is new. (Indeed, the word 'news' is used as a synonymous of 'information' in this meaning.) We may think of an information as increasing the recipient's knowledge inasmuch as he/she is able and willing to memorize it. The main feature of information, the *meaning* conveyed by the message, escapes any measurement and can be thought of as intrinsically qualitative. What can be measured, however, is how infrequent or unexpected it is: for instance, 'a man bites a dog' is much more informative than 'a dog bites a man' which refers

G. Battail, *Information and Life,* DOI 10.1007/978-94-007-7040-9_2,

to a more frequent event. If the probability of the reported event can be assessed, its unexpectedness can be measured by its improbability. This is how information is quantitatively measured according to Shannon (1948) (see Sect. 4.2.1 below).

A characteristic feature of an information in this meaning is that the reported event and its circumstances are only perceived by the recipient through the agency of a message conveyed by a channel. In the important case where this message consists of a sequence of symbols[1], like a spoken or written text, we will refer to the information as 'symbolic'. Then, the message consists of a sequence of symbols which can evoke the reported event in the recipient's mind, although there is no causal relation between the message and the event in the physical world. This is possible only insofar as the recipient can understand this message, i.e., provided the source and the recipient share a common linguistic system which consists of a set of conventional rules. The message has then a meaning within, and only within, this system. At the recipient's end, a communication thus involves two successive steps: the message has first to be received; then, using linguistic rules enables perceiving its intended meaning. The first step is performed by processing the channel output and results in making the message available to the recipient. The second step involves using the linguistic rules obeyed by the source for recovering the intended meaning, given the received message. Clearly, the first step is a mandatory prerequisite to the second one. It is easily overlooked as seemingly trivial, but communication engineers know by experience that it is far from being so. It should be emphasized that these two steps concern entirely different functions. Dealing separately with such unrelated problems is not only possible, but it is a methodological necessity. The competence of information theory is restricted to the first step of delivering the message to its intended recipient. We refer to this function as 'literal communication'. As not involved at this step, the recipient is no longer necessarily a living being (especially a human), but may be a machine as well. In any way, the literal communication between the source and the destination can entirely ignore linguistic and semantic aspects, hence the possible meaning of the message does not matter for it.

Still another fundamental property of an information is that it is non-autonomous. It is necessarily embodied within some physical medium which can be made of several different substances, devices or waves, and assume several forms. For instance a text can be written or spoken. The written text is made of a succession of visible marks of conventional shapes on some sheet of paper or computer screen, while the spoken text is represented by a succession in time of acoustic waveforms by the agency of which a listener can perceive the text. Admittedly, the spoken text has specific features like pitch, timbre, rhythm, intonation, accent, ... that the written one lacks, but as an information the text itself is common to both. In the technical field, a symbolic sequence can be recorded in the form of a binary sequence in the memory of a computer, then read and broadcast, for instance, as a 4-phase modulated electromagnetic wave. Then, both the alphabet size and the very nature of the physical medium which bears an information can be modified: the computer memory and the electromagnetic wave then bear the same information. In the absence of

[1] A symbol is an element of some given finite set of distinct objects, referred to as an alphabet.

any physical medium, however, an information cannot have any interaction with any material device or observer. Even our most abstract thoughts manifest themselves by the activity of neurons in our brain. We may say that no information exists unless it is borne by a physical medium, this remark being used for explicating what is meant here by the *existence* of an information. In what follows we refer to the physical bearer of an information as its *support*.

This way of defining existence meets a fundamental concept of Buddhist philosophy, according to Matthieu Ricard and Xuan Thuan Trinh (Trinh 2011, Ricard and Trinh 2002). It also complies with Carlo Rovelli's relational point of view on physics (Rovelli 2004) which has roots in ancient Eastern philosophy, too. Ref. (Ricard and Trinh 2002, p. 46) contains indeed the following quotation from the Indian philosopher Nagarjuna, who lived in the second century: 'Phenomena draw their nature from a mutual dependence and are nothing by themselves.'

2.2 Features of Information as a Scientific Entity

The few simple remarks above suffice for endowing the concept of information with the status of a scientific entity, and especially for founding a quantitative theory of information. As stated above, only finite alphabets are contemplated. This restriction entails that only discrete information is considered in this book although the theory has been fruitfully extended to continuous information. We prefer not to deal with this extension, which is not needed in the examples considered in this book, because it involves mathematical difficulties and some of its results are weaker than the homologous ones of the discrete case. Other applications to biology or physics would nevertheless need the extension of information to continuous random variables.

Dissociating information and meaning The first remark above enabled us to distinguish literal and semantic communication, and to state they are unrelated. Then, literal communication constitutes in itself a function which can be dealt with independently of any semantic consideration. In what follows, the word 'information' will be restricted to designate what can be *literally* communicated, and the mathematical theory of literal communication will be referred to as *information theory*.

This founding divide has been clearly stated by Shannon. He did not deny that information has something to do with meaning, of course, but he realized that communication engineering is entirely foreign to semantics. He wrote in the very first page of his seminal paper (Shannon 1948):

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design. (Shannon's italics.)

'Communication' is intended here in its *engineering* meaning of literal communication, foreign to the *philosophical* problems of semantics. Information processing devices and quantitative information measures are relevant only to the former. Indeed, a messenger has not to know about the content of the message he or she carries, and the same is true for communication machines, all the more semantics is intrinsically foreign to them. Ignoring semantics simply made information theory and its innumerable engineering applications possible, enabling the use of *mathematical* means for dealing with literal communication.

An information needs a physical medium but does not otherwise depend on it Indeed, an information has no existence (in the above meaning) unless it is borne by a physical medium, but a given information can be borne by any medium. In other words, an information is invariant with respect to the medium which bears it. It is thus an entity in itself. A medium is needed for embodying an information, but has no influence on it beyond securing its existence.

That an information does not exist unless it is borne by some physical medium, one of our basic postulates, sharply contradicts the common perception of informations (or of ideas) as purely abstract entities. For instance, no physical medium is as large as to memorize the immense knowledge that Laplace's omniscient demon is assumed to possess (see Sect. 6.3 below), so this demon cannot be but a pure spirit, hence foreign to the physical world. Stating that 'Information is physical', Landauer has to be credited for having challenged this opinion[2] (Landauer 1996). We are far from endorsing Landauer's statement, however, deeming instead that information needs to be physically inscribed. This statement is quite different from Landauer's but departs from idealism and similarly intends to anchor information within the physical world. We criticize Landauer's statement in Sect. 2.3 below.

Information is not conserved and can be shared Since its very existence depends on the medium which bears it, an information is *annihilated* if this medium is destroyed or incurs any change which alters the information it bears (altering an information, according to our viewpoint, replaces it by another one). Thus, contrary to many entities met in physics, information is not conserved. On the other hand, an information written on some medium can be copied on another one without being lost. An information can be copied several times so it can *proliferate*, meaning that *the same* information is borne by *several* different supports in increasing number. Proliferation does not mean any increase of information quantity, but that an information can be simultaneously borne by several supports. It is only when *differences* among the set of copies of a same information are created that the quantity of information it bears increases.

The ability of information to proliferate entails in the living world the ability of individuals to proliferate, which is a characteristic attribute of life. It is only because we deal with information as an *abstract* entity that we can reach this conclusion.

[2] In the information-theoretic literature. Physicists do not even ask the question whether information is a physical entity but deal with it as such, following Schrödinger and Brillouin.

Dealing with information as physical as did Landauer has not this consequence, hence is not adequate for biological applications.

An information as an equivalence class We notice that an information is invariant with respect to:

a) the physical nature of the medium which bears it (e.g., computer memory, acoustic or electromagnetic wave, sheet of paper, ...);
b) the alphabet size, which can be arbitrarily chosen for its practical convenience. For instance the binary alphabet is very convenient in operations performed by computers, but quite inconvenient for humans who prefer alphabets of larger size like the Latin one which are much better fitted to their perceptive performance. It is why programming languages contain instructions written in the Latin alphabet which are later converted into sequences of binary digits which control the machine;
c) for a given alphabet size, the possible transformation of a sequence into another equivalent one, related to the original sequence by a one-to-one correspondence according to some *encoding rule*.

Properties of invariance are, by essence, cumulative. The invariance stated by (a) and (b) is just a rewording of the above remarks. That stated by (c) is the most important but it is also the less obvious. It is why we will lay emphasis on it, especially in its form referred to as channel coding, in Sect. 3.4 and in Chap. 5 below. An entity defined as the set of elements which are invariant with respect to a transformation is referred to as *an equivalence class*. (Defining an equivalence class is a standard means for creating a mathematical object.)

Then *an* information is an equivalence class of sequences with respect to transformations a), b) and c). Dealing with an equivalence class implies designating a *representative* of it, which may be any of its elements. The most convenient representative of an information is the *shortest binary sequence* which belongs to its equivalence class, to be referred to as its *information message* (the physical medium needs not be specified as irrelevant to information theory, which by essence is mathematical). The symbols of the information message are necessarily independent[3] because, if they were not, source coding could transform it into a shorter one (see Sect. 4.3 below).

It should be kept in mind that an information thus defined as an equivalence class is quite an abstract entity, all the more it obviously contains infinitely many elements. As a collective object, a symbolic information may by no means be reduced, or likened, to a single sequence. Such a sequence may only act as its representative. Defining as we did an information as an equivalence class is a necessary consequence of the basic fact that an information must be *physically inscribed*. The multiplicity of possible physical supports of a same information suffices to show that it must be dealt with as an equivalence class. However, the necessity of encodings of different

[3] In any possible meaning of the word: there should not be a causal relation between them and, if they are random, they should be mutually independent in the probabilistic meaning of the word, i.e., their joint probability should be the product of their individual probabilities.

kinds which generally modify the length of a message is a still stronger incentive to do so. It is thus practical necessities of handling information which demand such an abstract definition, which is by no means gratuitous.

An information cannot be reduced to a number of any kind. It is an entity in itself, which meets Barbieri's concept of *nominable entity* (Barbieri 2007) (see Sect. 2.4.3 below). However, numbers can be associated with it. First of all and more important, the information message, its representative, can be interpreted as a natural number expressed in the binary numeration system and this may be useful, say, for classification purpose. Moreover, information theory enables associating a quantitative measure with an information, and it turns out that this measure equals the length of its information message. An information should by no means be confused with its measure but a difficulty arises as regards the vocabulary because the word 'information' is often used in papers dealing with information theory as an abridgement for 'information quantity'. Due to the dissociation of information and semantics, no ambiguity results in the engineering literature. When trying as we do here to make explicit the relationship of information with objects foreign to communication engineering, however, the reader must be warned against this confusion. We try to avoid it, sometime at the expense of rather lengthy periphrases.

Being defined as equivalence classes and not being numbers, informations cannot be ordered although their quantitative measures can be. Each information is an abstract object which has no other property than its *uniqueness*. No topology thus exists among informations, and the information quantity which measures an information is a mere attribute of it. As regards sequences, the Hamming metric to be defined in Sect. 2.4.4 and used later defines a topology among the sequences which represent informations, not among informations themselves.

2.3 Comments on the Definitions of Information

Shannon's approach in (Shannon 1948) was purely empirical. He defined an information *quantity* (see Sect. 4.2.1) but did not attempt to define information as an entity. It is such a definition which is proposed in the previous section, and the reader should be warned that this definition is not the only possible one. Indeed, defining information has not been necessary for developing information theory and was not needed for its engineering applications. We attempt here to define information in order to help applying it to objects foreign to communication engineering, i.e., so as to explicate its relationship with semantics (in a very broad sense).

We stated above that an information is *physically inscribed*. Let us emphasize how this is different from Landauer's statement, who wrote that 'information is physical' (Landauer 1996). He arrived at this conclusion by studying the physics of objects which can bear a binary digit, i.e., of two-state machines. But why should the physical properties of the support of an information be likened to that of information itself? Our statement that an information is physically inscribed leads to very different conclusions. For us, physical objects bear an information to which we attribute properties which have no physical counterpart, the most specific one being the possibility of its copy on a new support while keeping it on the initial one. Another very important property of information is that coding processes can transform a given sequence into another strictly equivalent one but having a different length. These two sequences obviously bear the same information, so an information cannot be likened to a single sequence, and still less to the physical support of such a sequence. The discrepancy between the status of information according to Landauer and to us is actually of capital importance, and it is only our information concept which will enable the conclusions of Chap. 10 regarding the relationship of the living and inanimate worlds.

We define an information, abstractly, as an *equivalence class* with respect to the possible supports which can bear symbols. Moreover, we consider as equivalent symbolic sequences deriving from each other by coding of any kind, i.e., by abstract transformations. We think that an information, instead of being itself a physical entity, has merely a mandatory relationship with physics due to the necessary physical inscription of a representative of it as an equivalence class. This will be examined in Sect. 3.1 and 3.2. The main parameter expressing the dependency of information on the physical world is the signal-to-noise ratio which determines according to Eq. (3.10) a symbol error probability. Symbol errors do not directly affect an information when it is represented by a word of a redundant code as efficient as to ensure its conservation. Instead of dealing with information as a physical entity, we propose in Sect. 6.3 to consider information as a fundamental entity from which we can derive the physical entropy, and not the other way round. Doing so is the exact contrary of what Schrödinger and Brillouin did (Sect. 6.3.4). Endowing information with the status of a fundamental entity will be especially useful in biology, as we shall see in the second part. It is the abstract definition of information that we propose, and only it, which can account for the specific properties of life. Using this definition entails that information appears as bridging the abstract and the concrete, as shown in Sect. 6.4. Only this definition can account for the fact that the abstract content of a symbolic sequence instructs the assembly and the maintenance of concrete objects as does a genome. Far from expressing a purely philosophical disagreement, opposing the statement 'information is abstract' to Landauer's 'information is physical' is of fundamental importance for our project of refounding biology on the science of information.

The confusion of an information with its necessary support, which reifies information and hence denies its abstract facet, is not limited to Landauer's work. Most physicists implicitly accept it without a serious examination of the issue. I am moreover afraid that the comparatively new discipline of quantum information theory which intends to integrate quantum physics into information theory relies on the same confusion (for instance, the acronym 'qubit' was coined for designating a physical system, while 'bit' designates the quantitative unit of information in conventional information theory). I am even reluctant as regards this very approach, and it is why this book ignores quantum information theory. On the contrary, I wonder if quantum physics could *derive from* information theory, although no attempt to answer this question is made in the present book.

2.4 An Information as a Nominable Entity

2.4.1 Naming and Counting

Any human society is made of individuals, each of them is *unique*. Similarly, we are surrounded with objects which can be uniquely identified as possessing some distinctive properties. We refer to objects or beings which can be unambiguously identified as *singular*. *Naming* a singular object means associating with it a vocal or written label which unambiguously designates, evokes or represents it. This label or tag is a sequence of a finite number of signs which belong to some finite repertoire given once and for all. Naming is an act of language, hence specific to the human culture. The first naming systems were probably vocal, or some combination of vocal signs and gestures. The signs of the repertoire should be mutually distinct but they are otherwise arbitrary. We mostly restrict ourselves in the sequel to *written* texts, i.e., to sequences of readable signs (often intended to represent vocal signs, or phonemes). Then, the repertoire of signs is referred to as the *alphabet* and its elements as *letters*.

A wide variety of objects can be named. They may be living beings, or singular objects of the physical world, or a set of physical objects which share some common specific property within an equivalence class. Relations between objects or sets of objects can be named, as well as these objects or sets. Sets of objects and relations of any kind belong to the abstract world. As perceived by the human consciousness, mental objects too can be named.

Among abstract objects, certain equivalence classes referred to as *numbers* soon needed to be specifically represented by symbol sequences in order to incur a kind of processing referred to as *computation*. *Cardinal* numbers just tell *how many* objects of some kind are present in some given set and *ordinal* numbers tell the place *where* given objects of some given ordered sequence are located. Letters, i.e., the same signs as used for transcribing phoneme sequences, have sometimes been used for denoting numbers when combined according to rules specific to this purpose, as for instance in Roman numeration. Such representations of numbers were rather cumbersome and their use for computing was quite complicated. Using symbols specifically intended to represent numbers, the Arabic *digits*, much better fits the needs of computation. Together with the modern numeration system, it enabled performing computation by simple machines as that invented in 1652 by Blaise Pascal (then 19-year old) as well as by nowadays computers. The representation of numbers by numeration systems will be considered in more detail later (Sect. 2.4.2). The choice of the base of a numeration system is just a matter of convention: it should be such that its digits are conveniently distinguished. We inherited the base ten from the Greco-Roman antiquity and, in accordance with this chosen base, the Arabic digits of our numeration system are ten. The bases twenty and sixty have been used in certain cultures, however, and we still use the base sixty for measuring certain time intervals (hours, minutes, seconds), a legacy of the Babylonians. Humans easily distinguish signs which belong to alphabets of this order of magnitude. Computers generally use much smaller bases, especially the simplest possible one, 2, although calculation circuits using the base 3 were implemented in early Sovietic computers.

We have thus now in Western culture two main sets of written symbols: the letters of, say, the Latin alphabet, which properly combined in sequences represent the words of a language; and the set of signs, referred to as digits, which represent numbers. There is however no intrinsic difference between letters as used for writing texts in some language and digits used for denoting numbers and computing, since both are elements of an arbitrary finite set of symbols. What differentiates written representations of words and of numbers by sequences of signs is merely that they are interpreted according to different rules. A written word and a written number may thus be considered both as sequences of symbols of some alphabet; moreover, a same alphabet can be used for both. This alphabet can be assumed to be the simplest possible, i.e., binary, without loss of generality. Sequences of binary digits are used in computers for representing both texts and numbers. We use from now on the acronym ‘bit’ for *binary digit*.

As an example, the letters of the Latin alphabet are currently represented in computers according to the American Standard Code for Information Interchange (ASCII). Each letter is denoted by a 7-bit word according to a one-to-one correspondence: a lower case letter is represented by ‘11’ followed by the 5-bit sequence which represents its rank in the alphabet according to the binary natural numeration. For instance, 1100001 denotes ‘a’, 1100010 denotes ‘b’, etc. Capital letters use the prefix ‘10’ instead of ‘11’, so 1000001 denotes ‘A’, 1000010 denotes ‘B’, etc. Another representation of the Latin letters uses the 8-bit words which result from appending to a 7-bit word as previously defined a single bit such that the total number of ‘1’s in the word is even: then 11000011 denotes ‘a’, 10000010 denotes ‘A’, 11000101 denotes ‘b’, 11000110 denotes ‘c’, etc. Appending this eighth bit provides a rudimentary means of error control: if an error affects a single bit in the 8-bit word, the number of ‘1’s becomes odd so counting the ‘1’s in each word enables detecting that an error has affected a single symbol, but not correcting it. More sophisticated means, using longer words, can result in locating the error in the word and correcting it.

The previous remark was concerned with the problem of representing the letters of an alphabet (hence words of a language combining several letters) by means of digits. At variance with such words, which are semantically related with outer objects, numbers have intrinsic properties and are endowed with structures of their own. Their study constitutes an important part of the mathematical science. The representation of numbers of any kind relies on the structure of the most basic ones, referred to as the ‘natural integers’. Let us now have a look at it.

2.4.2 Defining and Representing Natural Integers

The mathematical definition of a natural integer is given in many textbooks. It is most often defined as an element of a set $\mathbb{N}$ which satisfies Peano’s axioms, namely:

1. $\mathbb{N}$ contains a particular element named ‘one’ and denoted by 1.
2. For any element a of $\mathbb{N}$, there exists in $\mathbb{N}$ an element b, referred to as the successor of a, denoted by $b = a'$. Then a is said the predecessor of b.

3. Any element of $\mathbb{N}$ has a predecessor, except 1 which has none.
4. Two natural integers are equal if, and only if, their successors are equal.
5. $\mathbb{N}$ obeys the axiom of recursivity, namely:
   (a) If a property stated in terms of some integer n is true for the number 1;
   (b) If it can be proved that, if this property is true for any integer m larger than 1, then it is also true for the successor $m' = m + 1$ of m (an integer b is said to be larger than an integer a, $b > a$, if b belongs to the successors of a; the 'successors' of a should be understood here as a', the successor of a', the successor of this successor, etc.);
   (c) Then this property is true for any integer larger than or equal to 1.

Two operations, addition and multiplication, are defined on natural integers: addition (denoted by +) defined as $a+1 = a'$ and $(a+b)' = a+b'$, and multiplication (denoted here by $*$) defined as $a * 1 = a$ and $(a * b) + a = a * b'$. These operations are associative and commutative. Multiplication is distributive with respect to addition, i.e., $(a + b) * c = a * c + b * c$.

Less formally, Henri Poincaré assumes that the operation $x + 1$, which consists of adding the number 1 to a given number x, is firstly defined and he notices that this definition, *whatever it is*, does not play any role in the reasonings to follow (Poincaré 1902). He defines the operation $x+a$, which consists of adding the number a to a given number x, assuming that the operation $x + (a - 1)$ has been defined. Then, the operation $x + a$ is recursively defined by the equality

$$x + a = [x + (a - 1)] + 1. \tag{2.1}$$

In other words, we know what $x + a$ means when we know what $x + (a - 1)$ means, and since we already know the meaning of $x + 1$, it is possible to successively define $x + 2$, $x + 3$, etc.

Being *recursive*, the definition using Eq. (2.1) actually contains infinitely many distinct definitions, each of them becoming meaningful only when the meaning of the preceding one is known. Having thus defined the addition of integers, Poincaré recursively shows that it is associative, i.e., $a + (b + c) = (a + b) + c$ for any integers a, b and c, and commutative, i.e., $a + b = b + a$. He then defines the multiplication of integers by the equalities

$$a * 1 = a \tag{2.2}$$

and

$$a * b = [a * (b - 1)] + a. \tag{2.3}$$

Once $a * 1$ has been defined by Eq. (2.2) and (2.3) enables successively defining $a * 2$, $a * 3$, etc. Poincaré also recursively establishes the properties of multiplication, showing it is distributive with respect to addition, i.e., $(a + b) * c = (a * c) + (b * c)$, and commutative, i.e., $a * b = b * a$, for any integers a, b and c.

If we ignore some subtleties of Peano's derivation which are motivated by the exigence of mathematical rigour, we may simply summarize how the natural integers

Springer

http://www.springer.com/978-94-007-7039-3

Information and Life
Battail, G.
2014, XIV, 260 p., Hardcover
ISBN: 978-94-007-7039-3