STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Bruno JACOBFEUERBORN
Deutsche Telekom AG, Germany

REFLECTIONS ON DATA, INFORMATION AND KNOWLEDGE

Abstract. Data, information and knowledge are fundamental notions of information and computer science. The eternal discussion on whether knowledge is objective, as adherers of Plato claim, or it is constructed is a thrilling philosophical exercise. We present our approach to defining data and information, and a way to define and making use of knowledge that circumvents the above dilemma.

Keywords: data, information, knowledge

REFLEKSJE NA TEMAT DANYCH, INFORMACJI I WIEDZY

Streszczenie. Dane, informacja i wiedza są podstawowymi pojęciami w nauce o informacji i informatyce. Odwieczna dyskusja na temat tego, czy wiedza jest obiektywna czy konstruowana, jest pasjonującym zajęciem filozoficznym. W artykule zaproponowano własne definicje danych, informacji i wiedzy w sposób, który omija powyższy dylemat.

Słowa kluczowe: dane, informacja, wiedza

1. Introduction

In this paper we define these three basic notions that determine the foundations of information science and computer science, namely the notions of data, information and knowledge. We shall conspicuously show that data does not equate to information, and that information does not equate to knowledge. It was none other than Albert Einstein who said: "*Information is not knowledge*". The American Nobel Prize poet T. S. Eliot went even further in his poem "The Rock" where he wrote: "*Where is the knowledge we have lost in information?*" We shall come back to Einstein's insightful assertion and Eliot's concern. Already now we signal that our definitions of data and information, and *ipso facto* knowledge, broadly borrow from linguistics and logic, and perhaps to a surprise of the reader from psychology and sociology.

2. Data

The very term *data* poses some grammatical problem since if we strictly obey grammatical rules it is a plural noun taking a plural verb, whose designatum is a collection of single objects dubbed *datum*, but one can observe that in contemporary English there is a universal tendency to use this term with singular and plural verbs depending on the speaker preferences and on how the narrative flows. The Oxford English Dictionary reads: "*In Latin, data is the plural of datum and, historically and in specialized scientific fields, it is also treated as a plural in English, taking a plural verb, as in the data were collected and classified. In modern non-scientific use, however, despite the complaints of traditionalists, it is often not treated as a plural. Instead, it is treated as a mass noun, similar to a word like information, which cannot normally have a plural and which takes a singular verb. Sentences such as data was (as well as data were) collected over a number of years are now widely accepted in standard English.*" (OED, Data). In our discourse we shall adopt a liberal approach in which both usages are allowed, yet with a preference to use the word *data* as a singular mass noun.

In ordinary language data are items obtained as a result of experiments, measurements, surveys, investigations, and calculations. They may be single substantive objects such as numbers or words, or organized in form of sequences (vectors), tables, graphs, charts, and the like. They are typically associated with quantitative descriptions of things and phenomena. Without introducing an advanced formalism we define data as elements of a certain language while by a language we understand a couple (A, G), where A is an alphabet, a finite set composed of some arbitrary symbols or signs, and G stands for a grammar (syntax) that is a generative mechanism made up of a finite number of rules defined by the author of the language, which tell how to structure the symbols in legal patterns.

For a given language L = (A, G) *data* is defined as a finite pattern of symbols picked up from A and formed (generated) by means of and according to the rules belonging to G.

It has to be strongly emphasized that in this approach the notion of data is entirely semantic-free; there is no meaning assigned to the data. It is just a structure of symbols (signs) built up in a way that is determined by a given grammar. In other words, data does not bear any semantics; it is an abstract (yet physical) purely syntactical arrangement that can further on assume various meanings depending on circumstances. This said one could legitimately deduce that language L is also a semantic-free entity. It is the case, indeed. As opposed to the ordinary understanding of the word 'language', which automatically considers language as an interlinked couple of syntax and semantics, here language is devoid of any semantic features. While doing this we follow logicians for whom languages are entirely syntactical tools for constructing formulas. They clearly separate syntax from semantics to fulfill one of the objectives of modern logic, namely to make inference a kind of calculus, subject to automation. This approach is also particularly convenient and fruitful when it comes to manipulating data by means of computers. The basic paradigm of information technology is that data and meaning are separated from each other for the meaning is provided by the execution of a program that operates on data. Since information technology is an important subject to discussion and the main methodology that permeates this dissertation, we borrowed from logics what was there developed for defining formulas to define data. Now, let us illustrate our approach by a couple of examples.

Example 1. Let A = {**A**, **B**, **X**, **Z**, **Y**} be the alphabet we decided to adopt (note that comas do not belong to the alphabet), and as our grammar G = {R1, R2} we assume two rules R1 and R2 that allow one to construct data, namely: R1 – data is a sequence of exactly 5 characters over the alphabet; R2 – only sequences that start with **B** can be data, and R3 – only sequences complying with R1 and R2 are data. Now, we can say that the following strings:

BXZYB

BXXXX

BYBAX

are data whereas the next three sequences are not data:

XXZYB (R2 is abused since the string does not start with B)

BYBA (R1 is abused since the length of the string is 4 rather than 5 characters)

BXZYK (R1 is abused since ‘K’ does not belong to the alphabet).

Example 2. Let L = (A, G) where:

A = {*Ana, Tom, phoneNumber, birthDate, lives, London, is, her, 1982-06-12, if, then, and,* _} is the alphabet of 13 symbols (note that comas do not belong to the alphabet but the underscore ‘_’ is part of the alphabet, and that the first twelve symbols are composed of more than one character; this is perfectly legal as our definition of the alphabet given above does not constrain us in this regard).

G = {R1, R2, R3} is the grammar composed of the following rules:

R1 – data is a finite sequence of symbols over the alphabet

R2 – a finite string set up as follows: *data*. *data* is considered data, where *data* has to be a legal data of L (note an iterative character of this rule)

R3 – only sequences complying with R1 and R2 are data.

So, the legal six pieces of data are:

a) *Ana_lives_in_London*

b) *Ana_phoneNumber_is_1982-06-12*
c) *Ana_birthDate_is_1982-06-12*
d) *Ana_phoneNumber_is_1982-06-12. Ana_birthDate_is_1982-06-12*
e) *if_Ana_birthDate_is_1982-06-12_and_Ana_lives_in_London_then_ Tom_is_her_brother*
f) *phoneNumber_if_then_and_lives*

Note that the fourth data is a result of concatenating the second and the third data as allowed by rule R2. The following two strings are not data:

John_and_Ana_lives_in_london	(R1 is abused since symbol *John* does not belong to the alphabet)
Ana_lives_in_London	(R1 is abused since all the symbols do not belong to the alphabet; they are not written in italic).

Once more we remind that data do not carry any meaning. But while looking at the six data given in Example 2, someone, who understands English, might have a feeling that the first five data are intended to bear some meaning. In fact, data are often prepared in such a way they could suggest how to interpret them without referring to the existing interpretation scheme. Incidentally, let us remark that a person who does not understand English could not grasp any intended meaning, if any, of the first five data. This is an important note suggesting that those (humans or machines) who are supposed to interpret data have to be endowed with specific "qualifications" in relation with these data. We shall come back to this important matter in Chapter 3 when presenting the notion of *information content*.

Having studied our definition of data and the two examples one could think that this definition was coined to produce data in form of strings only, and then one might ask oneself whether this definition is general enough that it could also cover other patterns, e.g. matrices or chemical structures, or even much more sophisticated and demanding structures such as multimedia objects. We claim that it is feasible to define languages to generate such types of data; however, as this issue exceeds the frames of this paper we leave pondering on what such languages might look like to the reader.

3. Information

Interestingly enough, the term *information* and its derivatives sound alike in very many contemporary languages. This is a clear indication that this notion is universal and, therefore, one can expect that different cultures and people similarly understand its meaning. This is undoubtedly a fortunate situation given the essentialness and importance of this notion for practically all domains of human activities.

As we can learn from a comprehensive and detailed study on the origins and history of the term information (Capurro, 2003) it has a Latin origin: *informatio*. This Latin word was used by our ancestors in two contexts, which are the act of shaping the mind and the act of communicating something. The Oxford English Dictionary provides, inter alia, the following descriptions: "(a) Knowledge communicated concerning some particular fact, subject, or event; that of which one is apprised or told; intelligence, news; (b) The action or fact of imparting the knowledge of a fact or occurrence; communication of news; notification; (c) The shaping of the mind or character; communication of instructive knowledge; education, training; †advice (obs.). Now rare". It has here to be noted that this respected and honored dictionary defines information in terms of knowledge, which is exactly what we mentioned above as a common quotidian practice.

We, however, shall not share this habit, but we welcome Dictionary's general approach to defining information, in particular that it is "shaping of the mind or character", which straightforwardly corresponds to our intent for considering information as a triad of cognitive, affective and behavioral dimensions. By and large, we define *information* as a composite that amalgamates four major interlinked components, namely:

1. The *information object* that is the thing that actually carries information be it for instance paper, a magnetic or optical disk, a stone or a clay tablet. In other words, it is a medium where information is embedded in form of an inscription, recorded sounds or videos, or pictures or something else. The information objects can be volatile (e.g. air waves carrying words or tunes) or durable such as recorded optical disks or a written papyrus.
2. The *information content* that is the intended meaning, the communication, the message that the issuer of the information wanted to record on the information object. Noteworthy, in order to express the content one needs a means of expression, i.e. a language whose part is the format of the recorded content; the language complexity might vary from a simple one to a sophisticated one and depends on the subject domain and issuer's competences. Now, we can explicitly say that this language is exactly the language L for data description that we defined Chapter 2. Typically, one who records the content, the author or issuer, has already in mind, at least approximately, the target, the receiver of the information. This topic however lies beyond the present discussion; we neglect the question on whether the content actually or to what degree reflects the intent of its author and whether the target is able to pick it up and understand.
3. The *information issuer*. This notion includes three sub-elements that are as follows:
   - The *information author* can be a man or a machine that creates information content. If the author wants to externalize the information content, i.e. to make it available to

other people or machines s/he/it needs to transfer it from her/his mind or its internal state to an external medium and thereby to create an information object.

- The *information publisher* is the agent (person, device, organization) that publishes the information, which may require reformatting it from the original format and/or changing the medium. Obviously, it may happen that the information author and information publisher is the same agent.
- The *information disseminator* is the agent (person, device, organization) that takes care about the information dissemination in a given way and environment. In practice the information publisher and disseminator and even the author may be the same agent.

4. The *information value* that is relative to its receiver who values the information content depending on his/her/it subjective criteria. The value of information can be assessed in cognitive, affective or behavioral terms. Cognitive value is related to knowledge, affective value is related to emotions, and behavioral value is related to skills that allow one to execute given tasks. Measuring information value is difficult, especially when it comes to affective value. Here we really enter into uncharted waters. It is attributed to Albert Einstein that "*Not everything that is countable counts, and not everything that counts is countable*". Some suggestions in respect of measuring affective value are given in [5].

As some of the relationships linking the four constituents of information are of a subtle nature let us elaborate a bit on the aforementioned notions. Note that a piece of the same content may be valuable for one agent but valueless for another agent, just because it is of no use or is not relevant to the interest of the agent. Similarly, the content put on a certain medium may be considered worthless by its receiver, but when recorded on a different medium it can become desirable and valuable, and this is not because of the monetized value of the medium itself, rather it is because this other medium combined with content elicits an affective reaction in the receiver, who as a human being is susceptible to emotions. So, Marshal McLuhan was undeniably right when he uttered his seminal phrase: *The medium is the message*. Of course, what he meant was much more general than merely noting that the medium influences the valuation of information; he also claimed that the medium shapes the information; in a way it may even co-create information. Given our definition of information, there is yet something more that we can add to this famous assertion. It is not only a medium that matters; it is also important who is recognized as the author and/or publisher/disseminator. This thread will lead us to the definition of knowledge presented in Chapter 4.

Now, let us take a closer look at the information content that is a core component of the definition of information.

We define information content as data to which an interpretation was assigned. The interpretation provides us with the type of meaning that we want to be the semantics of our dataset. In other words, this is interpretation that bounds the data with a specific meaning. We thus say that information content is meaningful data. Formally, we have an interpretation function INT, a finite set of data D, and a finite set of pieces of information content I. Thus, *information content* on data d ∈ D is defined as the value of function INT(d), where:

INT : D → I

We accept that for certain data the interpretation may not provide any value, i.e. the data under this interpretation are meaningless. We denote such a case by the sign of ⊙, i.e. if data d is meaningless under interpretation INT, then INT(d) = ⊙. Note that by defining interpretation as a function we allow one to have different data with the same meaning. The following example provides an elucidation of the above definition.

Example 4. Let *Ana_phoneNumber_is_1982-06-12* be a piece of data borrowed from Example 2. Now, if we assume that the interpretation (semantics) that we have at our disposal assigns to this data the information content "Ana's phone number is 1982-06-12" we can formally mark it as follows (Note that to represent the information content we use the natural language, in this case it is English):

Ana's phone number is 1982-06-12 = INT(*Ana_phoneNumber_is_1982-06-12*)

By the same token the meaning of the data:

if_Ana_birthDate_is_1982-06-12_and_Ana_lives_in_London_then_Tom_is_her_brother

is "if Ana's birth date is 12 June 1982 and Ana lives in London then Tom is her brother", then we can write:

if Ana's birth date is 12 June 1982 and Ana lives in London then Tom is her brother	=	INT(*if_Ana_birthDate_is_1982-06-12_and_Ana_lives_in_ London_then_ Tom_ is_her_brother*)

Yet what about the data *phoneNumber_if_then_and_lives*? Here, in our interpretation the meaning of this data is non-existent, i.e. INT(*phoneNumber_if_then_and_lives*) = ⊙

The second case of Example 4 is of particular importance because the information content obtained as a result of applying the interpretation function is an if-then syllogism that without any doubt is an inferential structure. We can then expect that given a data set D along with the interpretation INT we can get a set INT(D) of information contents over which one will be able to perform various inferences, as we do during our colloquial or even more rigorous conversations, including scientific discussions when information is uttered, exchanged and created. While illustrating the mechanism data→interpretation→information content we focused on textual data. For any textual data or the data based on symbols and their structured collections we have no doubts that this mechanism works. For instance for the kind of data presented in Example 4 the interpretation can refer to logic and a natural language such as English. In order to design a data description language L one can make use of the first order predicate calculus whose structure can to a significant degree model a grammar of natural languages or even to make use of more advanced formal languages to depict various types of modalities such as defaults allowed in Reiter' Default Logic (Reiter, 1984) . Having said that now the question arises whether this mechanism can also be applied to other types of data, for instance multimedia data. Unfortunately, we do not have a workable proposal at hand and we have to leave this question open for further research, beyond the scope of this dissertation. Though one remark can be made, namely that a natural language can be used to express an interpretation associated to multimedia information objects. This is exactly what humans do while describing what they can see on a photo or while watching a move. Yet this is the final result of a complex interpretation process that as of now remains hidden behind the veil of the human's brain mystery.

Valuing information is a substantive and particularly difficult and complex topic. In this paper we do not discuss the whole spectrum of this problem, rather we refer the reader to the literature, in the first instance to the pioneering works by Claude Shannon and Warren Weaver [7], where information is treated from a purely technical point of view as an object (message) converted to electronic signals to be stored and transmitted as quickly as possible by means of available electronic channels with the least possible error rate. Their concern was how to ensure the efficiency and accuracy of the information transmission between the issuer and the receiver. In this situation an obvious gauge of information is the number of bits (signals) needed to encode the message. This gauge of course has nothing to do with the value of information as understood by humans, as it abstracts from the information content. The method of measuring information by counting the number of bits to represent it is of very little use from our standpoint; it is however a pretty good gauge to measure the volume of data.

Along with the engineering works on making transmission as accurate and reliable as possible Shannon and Weaver developed a Mathematical Theory of Communication that is based on statistics and probability theory, in which they assumed that the value of information is proportional to uncertainty, or in other words inversely proportional to predictability. We sympathize with such an assumption since it simply means that the less expected or the bigger surprise is the information content the higher is its cognitive value, and vice versa, the more predicted was the information content of the received message, the less cognitive value it includes. Note that this formulation assumes the subjectivity of information content valuation since the level of unexpectedness UNEX is relative to the information receiver. Yet, in spite of taking into account the subjective aspect of the information valuation for us this model for defining information content value is still not sufficient, as it does not include another vital subjective factor that impacts the information content value, namely the importance IMP of information. Unexpectedness and importance of information are not the only ones. Information receivers may take also into account the trustfulness TRU of the information source *s*. And again, certain pragmatically oriented receivers might be particularly interested in the usefulness USF of information content *i* and argue that usefulness impacts how they valuate it. By a similar token the value of information content i may depend on the context CONX determined by information contents of n information objects: $i', i'', \ldots i^{(n)}$ related to *i*. As a result we obtain the following relationship between the value V of information content *i* and factors that decide on this value:

$$V(i) \sim \mathrm{UNEX}(i),\ \mathrm{IMP}(i),\ \mathrm{TRU}(s),\ \mathrm{USF}(i),\ \mathrm{CONX}(i', i'', \ldots i^{(n)})$$

Needless to say that this is but a rough depiction, subject to further detailed elaboration and development to take into account other factors that might have an impact on the information valuation, including the ones referring to the behavioral aspect of information. A calculation formula is also needed to calculate the value of information content. All this lies beyond the assumed scope of this paper, but it needs to be signaled here that one vital issue is not often and comprehensively addressed in the literature on information, namely the asymmetry of information value as seen from the receiver viewpoint and the standpoint of the information issuer/publisher/distributor. This question is of special attention to information issuers, publishers, and distributors for whom information is tout court a commodity, subject to bringing profit, or to politicians, activists and all those whose mission or profession is to preach and convince their audiences. We conclude the discussion on information content valuation by expressing our conviction that there exists a common denominator, a ground to determine the cognitive, affective, and behavioral values. This common denominator is experience that embraces all three of them.

4. Knowledge

The attentive reader has probably noticed that information objects are carries of data, and in turn data are carriers of information content, and now when we tackle the question of knowledge we shall see that information is a carrier of knowledge. Such readers have also perceived that so far we have not discussed a crucial matter of logical value of information, neither in terms of binary logic (that allows two values only: true and false) nor multivalued logic. This omission was entirely purposeful since the question of validity of information content leads us to defining knowledge what is the subject of this subchapter. While doing so we abstract from the discussion on whether there is such a thing as absolute truth and whether knowledge can be absolute or it merely is a better or worse approximation of the truth. Knowledge is gained, or better to say, constructed through experiences, in science always subject to an ineffable effort of falsification.

The seminal sentence "*All men by nature desire to know*" opens Aristotle's "Metaphysics" [1] Having read it the immediate question that comes to mind is as follows: From where can men acquire knowledge? In the same opening paragraph Aristotle renders the answer: "*Experience is knowledge of individuals*". Individual experience gained every day along with an oral transmission of a generalized experience through generations was indeed the major source and vehicle of knowledge over centuries. The Enlightenment and scientific revolution that put emphasis on collecting raw real-world data, as a first stage of scientific investigation, have dramatically changed the situation. This is how natural sciences such as biology, chemistry, geology, physics and other disciplines have approached problems of their interest since the16th and 17th century. Francis Bacon's new methodology of science and knowledge, *empiricism*, that relayed on observation, collection of data, and experimenting, along with accepting induction as a legal inference method for scientific endeavors can be characterized as data-centric. Indeed, innate concepts, a priori assertions based on tradition, intuition or revelations could not be accepted as knowledge until they were verified and confirmed by rigorously organized experiments and the data the experiments yielded. Baconian science posits that theories that are meant to be the models of reality are derived from the analysis and generalization of the collected data and observations, or if the models are established as intellectual hypothesis, they must be verified through experiments producing data that in turn have to be examined and scrutinized [2]. This Baconian paradigm is presently applied not only in the scientific realm but also in a modified and less rigorous form in other domains such as marketing, politics or governance to learn social preferences and moods. Yet not so long ago, the trouble with the Baconian approach was that we did not have enough data to draw conclusions and build up models, or on the contrary, which is a more recent stage, there was too much data for a man or even a team to grasp it and discover patterns and regularities. Obviously the second case seems to be simpler to deal with, especially when large computers are available to crunch data, but paradoxically that was not the case until the last decade of the previous century when new opportunities came along owing to the increase in computer power, proliferation of the internet and proprietary networks, and emergence of new data analysis techniques such as data and text mining.

Over centuries, since the golden period of philosophy in ancient Greece there have been a lot of discussions and literature on knowledge. The theory of knowledge, called epistemology (from Greek ἐπιστήμη – epistēmē, "knowledge, science" + λόγος, "logos") has always been one of the most important and heavily cultivated branches of philosophy, concerned with the nature, scope and limits and boundaries of knowledge. It addresses such questions as: What is knowledge? How is knowledge acquired? What do people know? How do we know what we know? Do we discover or invent knowledge? Let us now systematize what is written above in this Chapter and attempt to establish a definition of knowledge. Of course, we realize there already exist many definitions of knowledge sometimes fairly contradictory one to another; incidentally, one of the most elegant and at the same time operational approach to knowledge states that: One who has knowledge of objects and their relations can classify objects, and vice versa, if one can classify, then one knows (has knowledge).

There are a number of definitions of knowledge. Our definition(s) of knowledge does not aspire to reconcile or compete with the existing ones. Its only ambition is to set up a pragmatic approach to understanding knowledge and being able to answer the question whether an assertion or a set of assertions is or is not knowledge. Let us consider a set of experiences, convictions, and/or beliefs, and even prejudices that fulfill the requirements specific for information objects, for instance they might be written or uttered sentences in a natural language. Thus, we have a set of information I at our disposal. Colloquially, *knowledge* K_C (casual knowledge) is understood as a subset of I, i.e.

$$K_C \subseteq I$$

Note that this definition does not require I to include exclusively true information (although it would be a much desirable case), and it even allows I to be contradictory (i.e. to include simultaneously statements such as ‘s’ and ‘not-s’). Moreover, it does not insist on explicit demonstration of how reasoning over the information content included in K_C goes. At the first glance such a definition looks like nonsense. But we human beings, at least a lion’s share of us, in many daily situations prove that this is exactly the case; we rarely check out the validity of various claims we stick to or express, and our reasoning does not always feature certainty; we are unconscious victims of diverse fallacies, patterns of poor and unreliable reasoning that appears to be sound, or we consciously and cynically exploit fallacies for winning arguments (which of course is exceptionable). Even more interesting is the fact that in one area our knowledge is consistent and non-contradictory because it is a rather narrow technical knowledge acquired from training, practice, and acknowledged authorities, and in another area it is just a colloquial popular knowledge, for instance about UFO, laden with inconsistencies. Incidentally, a relatively new discipline, dubbed *informal logic*, attempts to capture and understand the way people hold conversation and carry out arguments, and what kind of reasoning patterns they use for justifying their stances and claims. R. H. Johnson and J. A. Blair who are the founders of informal logic describe it as: “*A branch of logic whose task is to develop non-formal standards, criteria, procedures for the analysis, interpretation, evaluation, criticism and construction of argumentation in everyday discourse”* [4].

The loose approach to knowledge that is typical in everyday discourses or of people neglecting rhetoric and intellectual discipline cannot be accepted in philosophy and its daughter – science. But there is one interesting and significant difference between the two. It is not required from philosophers that the knowledge they offer to us in form of philosophical systems is true; however, we do require there are no contradictions within the systems, and the inference patterns are sound and visible. Science requires more: its knowledge has to be consistent (the logical value of scientific assertions is assumed to be true) and non-contradictory. Our approach to knowledge tilts to what is in this respect emblematic of science; yet, without all these rigors insisting on "objective truth". So, what does it look like?

In defining knowledge we are inspired by the notion of theory as it functions in logic [3], There is however one significant difference, i.e. theories in logic do not explicitly include meaning of formulas (however they are interested in their logical values), whereas our definition of knowledge explicitly contains the meaning of data (equivalent of formulas) that is provided by an interpretation. And in our approach to knowledge we add a crucial criterion, which is the validation of knowledge through an act of its accreditation. Formally, *knowledge* is the following quadruple:

$$K_F = \langle L, I_{INT}, SAX, RINF \rangle$$

where L is a data description language (as defined in subchapter 2.1); I_{INT} is a set of information content obtained by means of interpretation INT, where INT(D) = I_{INT}, and D is a set of data generated by L (as discussed in Chapter 2); SAX is a set of logical axioms (tautologies such as Law of Contradiction); and RINF is a set of inference rules (e.g. deduction, induction, inference by analogy). Of course, the kernel of this definition is the set I_{INT} including information contents, which provides the account of what we know about the subject matter depicted by our knowledge. Also important are inference rules since they decide on how reasoning over the information objects is carried out. Reasoning is particularly vital for it can be used for two purposes, i.e. for discovering new information occurring in the represented world yet not explicitly mentioned in the set I_{INT}, and for answering queries addressed to K_F, i.e. to our knowledge (answering is a two-step process: a query is considered a hypothesis, and then, one tries to prove the hypothesis; an illustration of this process is given in Example 5 below).

Having formally defined knowledge K_F we realize that this notion still suffers from a lack of warranty that the knowledge, especially I_{INT}, is certain so that one can relay on it and, if necessary to make use of it in one way or another. Here two cases are possible regarding the component I_{INT}:

a) the information content *i* is objectively verifiable; for instance it is a fact such as "Warsaw is a capital city of Poland" that can easily be checked out by referring to the Constitution of *Rzeczpospolita Polska*, or it is a statement that is true by sheer virtue of logic and facts it includes, e.g. "All men are mortal. Socrates is a man. Therefore, Socrates is mortal".

b) the information content *i* is claimed valid by an authoritative agency; examples of such agencies are science and communities of scientists, law makers, gurus and the like. There are two things we have to emphasis in this case:
   - we do not require the information content is true; the only requirement is that it has to be authorized (here we leave apart the question of the authoritative agency responsibility).
   - the decision on what is the authoritative agency that validates the information content is up to the possessor (receiver or creator) of the information content.

and also two cases regarding the component RINF:

a) the inference rule is sound by sheer virtue of logic, e.g. *modus ponens* or Robinson's resolution principle [6].
b) the inference rule is authorized and accepted to be used for carrying our inferences over the I set by an authoritative agency; examples of such agencies are administrators of knowledge base system, data and text miners, or scientists exploring big datasets generated by laboratory experiments.

The act of authorizing information content or an inference rule by an authoritative agency is dubbed *accreditation*. Let K_F = <L, I_{INT}, SAX, RINF> be a formal knowledge as defined above, and

$A_I : I \rightarrow \{0, 1\}$	be an information content accreditation function, which assumes value 0 when an information content does not obtain accreditation, and value 1 when it is objectively verifiable or obtains accreditation.
$ARINF : RINF \rightarrow \{0, 1\}$	be an inference rule accreditation function which assumes value 0 when an inference rule does not obtain accreditation, and value 1 when it is sound or obtains accreditation.

Eventually, we can define *accredited knowledge* K, or simply *knowledge* K as follows:

K = < L, **IA**, SAX, **RINFA**>

where: **IA**=$\{\boldsymbol{i} \mid \boldsymbol{i} \in I_{INT}$ and $A_I(\boldsymbol{i}) = 1\}$, and **RINFA**=$\{\boldsymbol{inf} \mid \boldsymbol{inf} \in RINF_F$ and $A_{INF}(\boldsymbol{inf}) = 1\}$.

This definition permits of individual or “collective subjectiveness”, an arbitrary look at what knowledge is. It reflects our conviction that although knowledge can include assertions that seem to be unconditionally, say objectively true, it is nevertheless constructed. For knowledge to become public its issuer has to detach it from its creator (who as already mentioned may be also its issuer) and submit it to a relevant community, the authoritative agency, for accreditation, and only aftermath this new knowledge can be integrated with the existing corps of knowledge.

5. A Final Remark

The eternal discussion on whether knowledge is objective, as adherers of Plato claim or it is constructed on the basis of data, experimentations, reasoning and building conceptual models, often formulated in the language of mathematics, is a thrilling philosophical exercise. Our approach to defining knowledge circumvents this dilemma as it allows one to include in the corpus of knowledge claims and theories coming from both sources, i.e. from Platonic approach and constructivism provided that they are accredited.

BIBLIOGRAPHY

1. Aristotle: Metaphysics. 350 B. C. E., http://classics.mit.edu/Aristotle/metaphysics.html (20 Feb. 2013).
2. Bacon F.: The New Organon or True Directions Concerning the Interpretation of Nature. 1620, http://www.constitution.org/bacon/nov_org.htm, (20 Feb. 2013).
3. Hurley P. J.: A Concise Introduction to Logic, 7th edition. Wadsworth Pub Co, 1999.
4. Johnson R. H., Blair J. A:. The Current State of Informal Logic. Informal Logic, 9(2–3), 1987, p. 147÷151.
5. Kaczmarek J.: Representation of Experience in Information Systems. Ph.D. Thesis in preparation. Institute of Information Science and book Studies, Warsaw 2013.
6. Robinson J. A.: A Machine-Oriented Logic Based on the Resolution Principle. Journal of the ACM, Vol. 12, Issue 1, 1965, p. 23÷41.
7. Shannon C. E., Weaver W.: The Mathematical Theory of Communication. Urbana, University of Illinois Press, 1949.

Submitted on the 16th January 2013.

Omówienie

Dane, informacja i wiedza są podstawowymi pojęciami w nauce o informacji i informatyce. Pojęcie danych nie wzbudza kontrowersji, natomiast pojęcia informacji i wiedzy, a zwłaszcza wiedzy są przedmiotem dyskusji i sporów. Odwieczna dyskusja na temat tego czy wiedza jest obiektywna czy konstruowana na podstawie danych, eksperymentów, wnioskowania i budowania modeli konceptualnych, często formułowanych w języku matematyki

jest pasjonującym zajęciem filozoficznym. W tym artykule proponujemy taki sposób definiowania i posługiwania się wiedzą, który omija ten dylemat.

Address

Bruno JACOBFEUERBORN: Deutsche Telekom AG, Landgrabenweg 151, Bonn, Germany, Bruno.Jacobfeuerborn@t-online.de.