AMO - Advanced Modeling and Optimization, Volume 16, Number 1, 2014

Conceptual framework for modelling knowledge acquisition via content-related Internet sources

Georgios Constantine Pentzaropoulos[1]

Mathematics and Information Technology Unit,
Department of Economics, University of Athens,
8 Pesmazoglou Street, 105 59 Athens, Greece.

Abstract

Knowledge acquisition is approached here from a theoretical/methodological point of view. We develop a conceptual framework for modelling knowledge codification and acquisition via interaction with the real-world, e.g. the Internet, elaborating on mechanisms for gaining *practical knowledge*. This effort requires crossing many fields, including computer science, artificial intelligence, epistemology, as well as the cognitive sciences. Reliable information is shown to reduce entropy and increase stability, thus allowing for more informed decisions.

Keywords**:** Actionable knowledge; codification and acquisition; decision rules; entropy; stability; reliability; system performance evaluation.
MSC Classification: 68A45, 68P20, 68M11, 68U35, 68T30.

1. Introduction

Man acquires knowledge via interaction with the external world. An intuition implies reference to memory is made, the required information in *codified* form is fetched, and a logical connection (synapse) follows. Man, then, enters a “state of reason”. Awareness of mind is linked to perception as the absence of observation would not allow a cognitive mind to capture objects and know them. Knowledge *acquisition* is a central subject in a number of diverse scientific fields including computer science, artificial intelligence, epistemology, and the cognitive sciences.

In this work we approach the problem of knowledge codification and acquisition from a theoretical/methodological standpoint, elaborating on the mechanisms which allow us to gain *practical knowledge.* This kind of knowledge, also called *actionable* by Argyris [1&2], is well correlated with skills thought necessary in today’s society. According to Floridi [3] knowledge and information are members of the same conceptual family. Since much of information today comes via the world wide web, it seems reasonable to concentrate on locations where such codified information is widely available. Such locations are collectively known as knowledge bases.

Knowledge bases (KBs) are today integral parts of large-scale computer networks. They contribute to our understanding of many contemporary subjects of interest as being key elements of a global information economy and society. KB efficiency relies largely on the capabilities of the underpinning telecommunications networks.

[1] *Associate Professor (ICT). Email: gcpent@econ.uoa.gr.*

AMO - Advanced Modeling and Optimization. ISSN: 1841-4311

The main features of these networks include an array of broadband technologies (fixed and wireless), access via intelligent interfaces, and better quality of service. Services presently being employed include: on-line searching in digital libraries, transfer/sharing of multimedia files, as well as collaborative work including decision-making across national borders.

Human-readable KBs, which are examined here, are widely distributed across interconnected servers supporting many users with varying educational backgrounds. Engines are used to mine data in KBs. Modern knowledge-based systems incorporate artificial intelligence techniques [4]. We examine this subject in subsequent sections. Many modern KBs can support, *inter alia*, web-learning and automatic information discovery within the remote parts of world wide web.

With regard to this article, we proceed as follows. First, we discuss fundamental concepts such as data, information, messaging, memory, and the relevant interfaces. Then, we go on to more complex subjects like codification, encryption, knowledge representation and reasoning, and finally knowledge utilization. We conclude with practical aspects within the framework of today's information society [5].

2. Binary data, information, and knowledge

Computer systems are employed for "information processing", which is their primary function in any given environment. In this section, we are interested in the nature and relationship between the following terms: *binary data; information; inference*; and, finally, *knowledge*. These terms are discussed within the the framework of artificial intelligence later in this article.

2.1. *Access to the real world*

Theory of knowledge, also known as Epistemology, is primarily concerned with the origin of knowledge, the place of experience in generating knowledge, and the place of reason in doing so. From this point of view, one might see two independent possibilities as to the way by which interaction between an observer/user and the "outside world" takes place.

P1. The observer might reason with abstract ideas only. Using his previous sense experience the observer can see a number of indistinguishable "possible worlds". Then, interaction with any of the worlds follows.

P2. The observer interacts with a "real world", the only world seen. However, this interaction often contains elements of *uncertainty;* hence, any knowledge gained here should be the subject of inspection.

There is also the problem of the assumptions that one frequently makes in order to claim knowledge. This problem is nicely illustrated by Klempner [6] in Chapter 4 of his book *The Possible World Machine* as follows: "What we call our 'knowledge' rests on a vast network of assumptions; assumptions whose truth or falsity would be impossible to check or prove in their entirety. We might feel safer making the more modest claim that all we really know is what each of us has learned directly through our senses. Everything else that we believe comes under the heading of the ‘best explanation’ of our sense experiences.”

Let it be noted that the above network of assumptions has a clear similarity to the well-known series of *“if ... then ... else”* statements often expressed as logical conditions in algorithmic programming languages.

The existence of a materialized “real world” is taken for granted in this study as such existence is compatible with the function of any information system; and KB systems are advanced information systems. Thus, we have to deal with uncertainty. The subject of uncertainty and its consequences are examined in more detail in later sections in the context of information entropy.

Mathematical models are necessary for constructing a framework to formally express any cause-effect relationships. Models that are *logic-oriented* are particularly useful here as they link information content to possible logical implications. There are also other kinds of models, e.g. probabilistic ones; and, of course, the current application finally determines which kind of model is more suitable.

Figure 1 shows a part of a possible “real world” of interest to an observer. Information transferred from this outside world to the observer, e.g. via a bit stream, might be thought to be converted into actual facts by means of logical inference. These facts can then be *transformed* into knowledge.

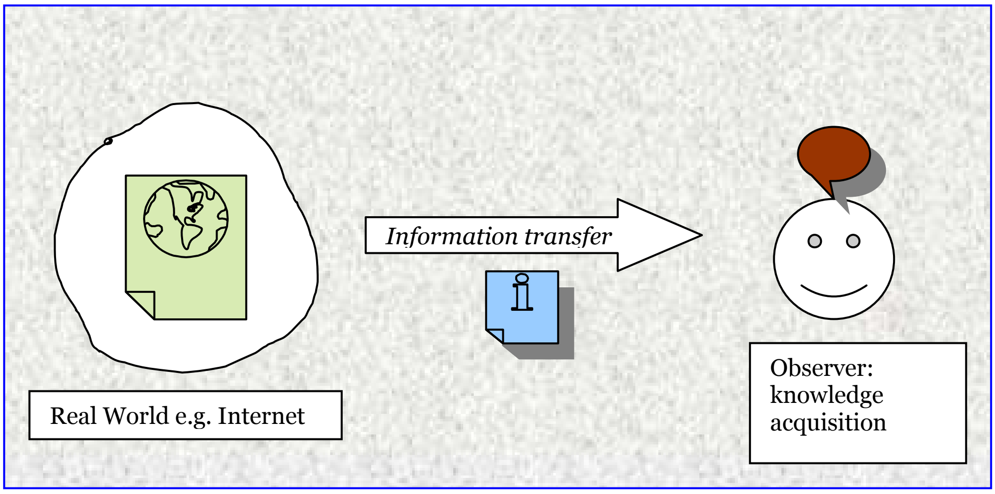

Figure 1: Knowledge acquisition via interaction with the external world.

For the purposes of this work, we consider the human brain as a *neural network* with complex circuitry and associated memory. At any time, the human brain is capable of selecting what it needs, store its content, and then retrieve it for later use. Memory is an important part of this learning process since it involves brain’s neuronal circuitry. Our aim is to examine how knowledge is acquired and managed in human memory. The analysis follows recent work of ours [7&8].

2.2. *Memory model*

Our starting point is the so-called *tabula rasa*, a well-known term which literally means some blank table. Therefore, at its initial state, the mind has no information. The problem now is this. The tabula is constantly exposed to the “outside world” of Figure 1 which keeps feeding it with new information. In the course of time the tabula is likely to become full unless we assume that it has infinite capacity. However, such an assumption cannot be realistic given the physical constraints that underpin human evolution. Therefore, memory can be very large, but not infinite, despite brain’s extensive neuronal circuitry.

In computer science, the concept of knowledge is usually defined as follows: *“Knowledge: the objects, concepts and relationships that are assumed to exist in some area of interest”.* Knowledge differs from data or information in the sense that new knowledge may be created from existing knowledge by logical inference. Information may, then, be thought of as the result of applying some kind of processing to some (raw) data, giving it meaning in a particular context. Thus, information can be assumed to be a prerequisite for knowledge. This assumption is also valid in the field of cognitive sciences. Following this discussion we formulate our first statement:

S1. *Data (binary representation)* [PLUS] *Processing (e.g. computer languages, data structures, algorithmic procedures)* $\Rightarrow$ *Information* $\Rightarrow$ *Knowledge.*

Computer scientists often make a distinction between main memory (MM) and virtual memory (VM). VM, unlike main memory, is always stable: it retains its content even when electrical power is switched off. This is because VM is a specific area of the system’s disk, which is stable by construction. Returning to our earlier discussion, we note that MM is like a *short-term* memory, a *tabula rasa*. MM by its construction contains recent information: thus, in relation to the field of neuroscience, it could be a model of the human brain’s hippocampus. This small part of the brain is responsible for holding information about recent events.

In contrast, VM could be seen as a *long-term* memory, e.g. a very large database containing billions of records. Such databases form the infrastructure of today’s digital libraries, and, of course, are sources of knowledge. Our second statement is, thus, as follows:

S2. *Short-term human memory* $\cong$ *random-access main memory (MM). Long-term human memory* $\cong$ *logically-structured virtual memory (VM).*

Again, in computer technology, *memory renewal* is based on the so-called criterion of temporal locality. The algorithm implementing this criterion is known as Least Recently Used (LRU). We recall that MM has finite capacity. Therefore, the continuous accumulation of pages would result in “overflow”, which is why memory renewal is necessary.

With respect to the human memory, LRU might be considered as a logical function embedded into the brain’s neural network circuitry. The processor itself could be thought of as being analogous to the entire circuitry, which contains all neurons and their synapses. Space in VM is purely logical and pages in VM have their own logical addresses. These addresses are then thought to be “mapped” to the physical addresses of main memory whenever page renewal is necessary. Following is our third statement:

S3. *Knowledge acquisition ≅ information codification, then, renewal in short-term memory, and mapping between MM and VM, aided by algorithmic functions based on the principle of temporal locality.*

3. Entropy of knowledge

Entropy of a system under examination can informally be defined as loss of order. Such a lack of order brings about uncertainty, which degrades system performance. In everyday life, information is commonly associated to a sense of order; hence, the need for most people to stay informed regarding their subject of interest.

In a recent article of ours [9] we concluded that *“information does not equate to knowledge”* which is in line with the earlier independent results, e.g. Greenfield [10]. Although information processing is not understanding, it can be argued that reliable data and careful processing leads to more informed decisions. However, practical evidence shows that sensory knowledge is not always reliable. Extracting knowledge through information processing, may sometimes be unreliable if the initial input contains errors. Such errors might be factual, as in case of distorted binary numbers, or purely logical, e.g. when a sequence of statements breaks down. Whenever that happens, knowledge becomes unstable.

Also in [9] above we examined the problem of *knowledge stability* both from an epistemic and systemic point of view. There, the final conclusion was that reliable information always leads to greater stability of knowledge. This effort requires reduction in uncertainty, which is considered as manifestation of a system’s entropy. Therefore, by reducing entropy, uncertainty is also reduced, and the system under examination becomes more stable.

3.1. *Information stability*

Entropy has its roots in physics, particularly in thermodynamics. In physics, the *entropy* of a system is defined as a measure of its intrinsic uncertainty. This notion has become familiar in our times thanks to the pioneering works of Claude Shannon and Norbert Wiener. Shannon introduced the term *“information entropy”* by means of his mathematical theory of communication. Wiener, the founder of Cybernetics, regarded information as the negative quantity of entropy and has shown that any increase in information will give greater stability, whether such information is communicated by humans or a machine [11]. Also, in a very different field, which is known as economics of information, Arrow [12] defines information as the reduction in system uncertainty.

This reduction in uncertainty, if examined through classical information theory, indicates some kind of "self-organization" within the system [13]. Knowledge-based systems can be seen as special cases of self-organized information systems [14].

Many real-world systems of our time, including academic/research environments, businesses as well as organizations, often have circulatory movement of information: such movement often takes the form of a *feedback process*, which contributes to increase in entropy. Since entropy, by its nature, distorts information content, its presence often leads to incomplete knowledge and away from the right decisions. This observation is evident in a number real-life situations involving interactions of the following forms: "humans ↔ humans", "humans ↔ machines", and in more advanced settings "machines ↔ machines". The above discussion naturally leads to the following statement:

S4. *Entropy minimization of a real-world system ⇒ Less uncertainty concerning information content ⇒ Reduced feedback processes ⇒ Greater knowledge stability ⇒ More informed decisions.*

The effect of entropy increase on information, and hence on the system's stability, can be described by means of the following figure:

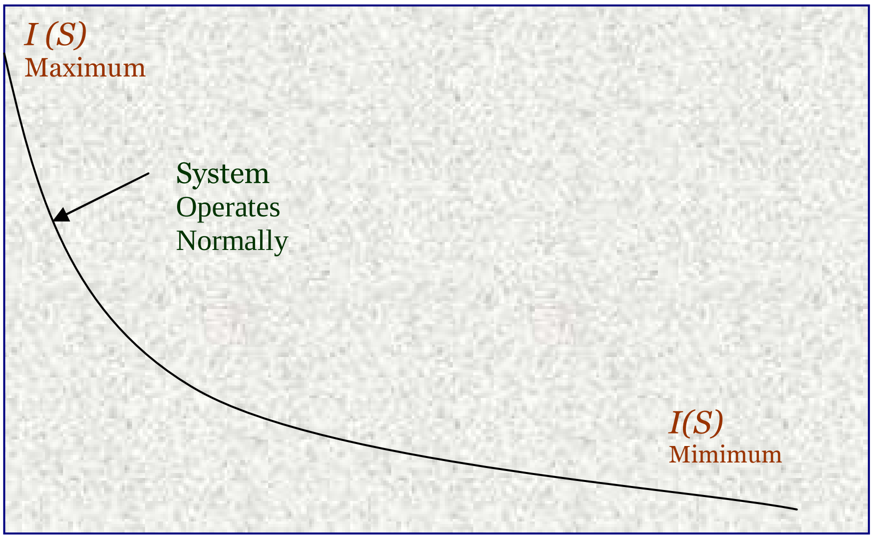

Figure 2: Information curve *I(S)* as system entropy *E(S)* → ∞.

Information often contains uncertainty, which comes from the methods of obtaining the raw data. Uncertainty can be characterized analytically, as will be shown later. Reliable information leads to greater stability, which can be achieved by some form of internal system organization. Thus, for now we may note that by reducing entropy, uncertainty is also reduced, the flow of information becomes normal, and the system operates near the the top left-hand region of Figure 2: the steady-state region.

3.2. *Acquired information*

According to Shannon's original work, information entropy is a measure associated with a set of possible system states. The following formula:

$$S = \sum_{i} p_i \, log \, p_i \qquad (1)$$

measures the average entropy of a system in the presence of probabilities $\{p_i\}$ that correspond to a series $\{i\}$ of system states.

When human communication with the real-world takes place, i.e. as suggested by Figure 1 given earlier, information transfer (e.g. in the form of bit stream) underpins this form of communication. Also, from the previous discussion, we may think of entropy as a measure of the amount of uncertainty about some expected outcome. Similarly, we may consider information as a measure in the reduction of uncertainty after observing a kind of logical connection or *hint* related to the same outcome. Having a hint helps the observer to form an opinion about the chance a specific outcome has to occur given a number of possible outcomes. Such an opinion can be described quantitatively in terms of probabilities.

Let us consider a system *S* and an associated hint *h*. Then, we express the entropy *E(S)* of system *S* as follows:

$$E(S) = -\sum_{s} p(s) \, log \, p(s) \qquad (2)$$

where *p(s)* is the probability of observing state *s*.

We can also express the information acquired with the help of hint *h* by classical arguments such as those found in information theory. This acquired information is quantitatively the reduction in uncertainty within the system.

We also express the conditional entropy *E(S|H)* of system *S*, after the observation of hint *h* as:

$$E(S|H) = -\sum_{h} p(h) \sum_{s} p(s|h) \, log \, p(s|h) \qquad (3)$$

where *p(h)* is the probability of observing hint *h* and *p(s|h)* is the conditional probability that system *S* is in state *s*, after the observation of hint *h*. *E(S|H)* measures the residual uncertainty of system *S*. Therefore, the difference:

$$A(S) = E(S) - E(S|H) \qquad (4)$$

shows the extent by which hint *h* reduces the uncertainty. *A(S)* is named here the *acquired information* and its measured value extends from zero - when the elements in the right-hand part of eq. (4) are equal - up to some positive number.

4. Knowledge from content-related sources

When considering communications and information networks like those depicted for example by Figure 1, the elements transmitted across are messages containing words, documents, symbols, graphs, and so on. These messages are typically measured by their information content. Such content can adequately be characterized by classical information theory arguments.

Messages may arrive at the receiving end in proper order, i.e. in the way it was indented by the sender, or their order may have changed due to random fluctuations. In the first case, messages retain their meaning while in the second case part of their meaning is lost. The latter case is especially troublesome since, in the mind of the receiver-observer, information content appears as a random aggregation of elements. Thus, in this case, we have a typical appearance of high entropy. As discussed in earlier sections, high entropy may lead to chaotic behaviour. Since we examine real-world systems for the purpose of improving their performance, the presence of high entropy is clearly undesirable and must therefore be minimized. Also, as suggested by Figure 2, as entropy tends to infinity information approaches zero.

Sometimes the system at hand becomes overloaded as messages arrive at a rate that its storage, such as buffer front-ends, cannot accommodate. System performance becomes problematic and, after reaching its peak point (plateau), begins to fall often quite sharply towards zero. This phenomenon is known as thrashing or performance collapse [15]. Figure 3 below shows the system's expected performance as a function of its workload measured by the number of received messages.

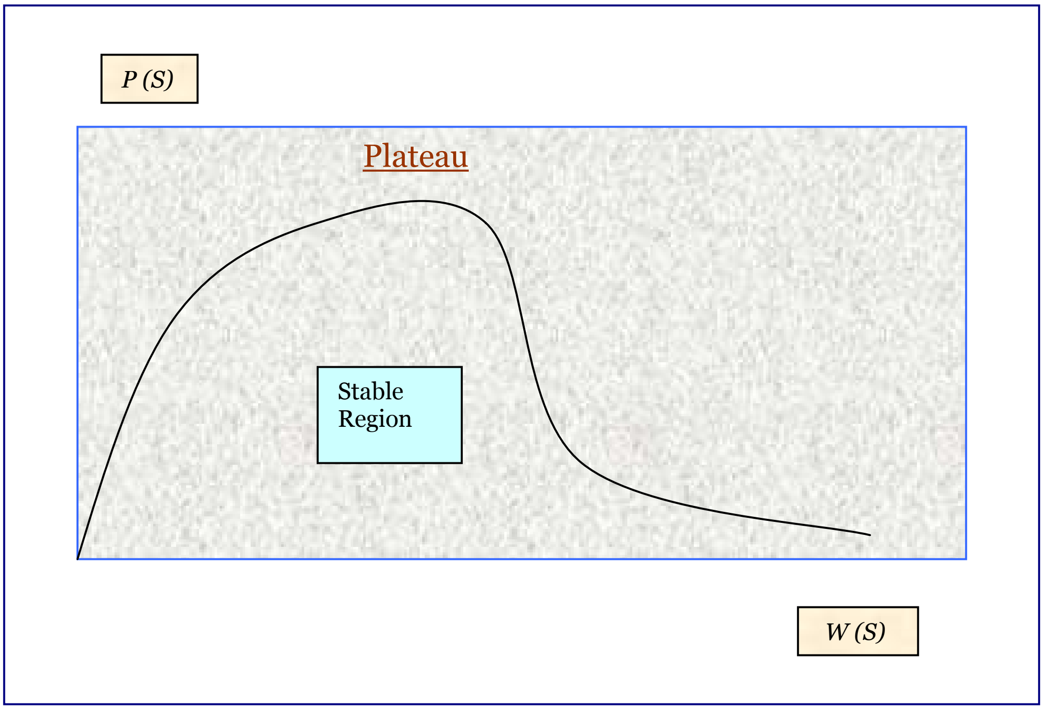

Figure 3: Thrashing curve as a function of system workload.

4.1. *From information to knowledge*

Entropy in the terms discussed above requires the existence of a human interface. Networks are able to provide plenty of information to the receiver-observer, but such information is often volatile. Error-correcting mechanisms are often used as means of reformatting messages in a way as to obtain their original meaning and retransmitting messages when they are lost. Message content along with information flow rates within a network can be used for constructing ordering schemes. Such schemes help towards information organization and retrieval at a later stage. Instruments that apply in this framework are:

(a) measurements of events;
(b) induction processes;
(c) deduction processes; and,
(d) mapping algorithms.

This brings us closer to knowledge representation, reasoning and, finally, acquisition. Principally, this is about the encoding of all propositions accepted by an observer. Artificial Intelligence (AI) practice is concerned with the construction of knowledge-based systems realized via suitable man-system interfaces [16]. A carefully designed system should have a number of desirable properties. Perhaps, the first property is the system's ability to perform some kind of a dialogue - via its interface - and then perform the following functions:

F1. Accept the observer's inputs and adjust itself accordingly.
F2. Implement qualitative reasoning based on experience.
F3. Invoke cognitive processes involving probabilities.
F4. Use combinations of possible events and decision paths.
F5. Convey knowledge in an effective manner.
F6. Operate a user-friendly fashion during a session.
F7. Implement the properties of integrity and safety.

Inference often involves backward-type chains, i.e. a mechanism involving at first some possible solution which must then be proven true or false. This interactive form of reasoning makes use of formal rules or examples stored in the knowledge base. Sometimes, heuristics are also used: these are informal rules associated with pieces of knowledge acquired by experience.

4.2. *Knowledge acquisition sessions*

The knowledge acquisition process is the most time-consuming task with a great deal of complexity. Because this process is quite unstructured, it cannot follow a form known *a priori*, e.g. a linear or an exponential form. After knowledge is conveyed, following a session, this knowledge will have to be stored and then presented into some suitable representation scheme.

Because knowledge acquired from sources of external environments is sometimes incomplete or even erroneous - as previously suggested - it follows that such knowledge should be tested for inconsistencies. This might give rise to further communication in order to clarify existing parts of information and/or elicit more information associated with the running session. The steps described above can be formulated as follows:

```
[S1]: Start of session.
[S2]: Transfer required information.
[S3]: Process information to form knowledge.
[S4]: Clarify knowledge and/or elicit more;
      then, go to [S2]; else, continue.
[S5]: Store knowledge in an appropriate format.
[S6]: Convert existing knowledge into representation;
      then, go to [S2]; else, continue.
[S7]: End of session.
```

4. Concluding remarks

The problem of knowledge acquisition has been approached, in the present work, from a theoretical/methodological point of view. We have developed a conceptual framework for modelling knowledge acquisition via interaction with the real-world, elaborating on the mechanisms which allow us to gain forms of *practical knowledge.* Knowledge acquisition lies at the intersection of diverse scientific fields including computer science, artificial intelligence, epistemology, and the cognitive sciences.

Knowledge bases are in our times integral parts of large-scale computer networks contributing to the understanding of many subjects within the realm of a new global information economy and society. The results obtained in the present work should be particularly useful in the context of this still evolving society. The medium is the Internet with its vast array of servers and extensive digital content [17].

Reliable information has been shown to lead to greater stability of knowledge. Stability requires internal organization as regards information, which can be achieved at the expense of uncertainty. The latter was described as a manifestation of entropy. Thus, by reducing entropy, uncertainty is also reduced, and the flow of information approaches steady state. In that state, knowledge is always stable and valuable.

Therefore, we are now able to express a composite final statement as follows:

S5: Final Statement. **(i)** *Binary representation of data and information processing are both prerequisites for acquiring knowledge from the real world, e.g. the Internet.* **(ii)** *The human brain may adequately be modelled as a complex neural network consisting of a short-term and a long-term memory.* **(iii)** *The process of knowledge acquisition requires information collection, codification, and representation realized by mapping algorithms.* **(iv)** *Entropy minimization implies less uncertainty and hence greater system stability: the end result is more informed decisions.*

Acknowledgements

This work has been partially financed by grant “elke/70/11/698” awarded by the Research Committee of the University of Athens, Greece.

References

[1] Argyris, Chris (1996a). Actionable knowledge: design causality in the service of consequential theory. *Journal of Applied Behavioural Science,* 32 (4), pp. 390-406.

[2] Argyris, Chris (1996b). Actionable knowledge: intent versus actuality. *Journal of Applied Behavioural Science,* 32 (4), pp. 441-444.

[3] Floridi, Luciano (2012). Semantic information and the network theory of account. *Synthese,* 184 (3), pp. 431-454. Springer.

[4] H. Yoshizumi, K. Hori & K. Aihara (2000). The dynamic construction of knowledge-based systems, in: C.T. Leondes (Ed.), *Knowledge-Based Systems: Techniques and Applications.* Academic Press, San Diego, California, pp. 560-604.

[5] Mansell, R. (2010). The life and times of the Information Society. Prometheus: Critical Studies in Innovation, 28 (2), pp. 165-186.

[6] Klempner, Geoffrey (2007). *The Possible World Machine*. Pathways Program A. Introduction to Philosophy.

[7] Pentzaropoulos, G.C. (2010). Knowledge acquisition as a memory renewal process. *Philosophy Pathways,* 159, Part I.

[8] Pentzaropoulos, G.C. (2011). Generating stable knowledge via reduction in entropy. *Philosophy Pathways,* 167, Part III.

[9] Pentzaropoulos, G.C. (2012). From data and information to actionable knowledge. *Philosophy for Business,* 75, Part III.

[10] Greenfield, Susan (2008). *ID - The Quest for Identity in the 21st Century.* Seminar on Tuesday, 9 December 2008, Thomas More Institute. Available at: http://thomasmoreinstitute.org.uk/papers/

[11] Wiener, Norbert (1950). *The Human Use of Human Beings.* The Riverside Press (Houghton Mifflin Co.).

[12] Arrow, K.J. (1984). *The Economics of Information.* Cambridge, Mass.: Belnap.

[13] Abramson, N. (1963). *Information Theory and Coding.* New York: McGraw-Hill Company.

[14] Ackoff, R. L. (1989). From Data to Wisdom. *Journal of Applied Systems Analysis.* Volume 16, pp. 3-9.

[15] Denning, P.J. & Buzen, J.P. (1978). Operational analysis of queueing networks. *ACM Computing Surveys* 10 (3), pp. 225-261.

[16] Russell, Stuart and Norvig, Peter (2010). Artificial Intelligence: A Modern Approach, 3rd Edition. Prentice Hall (Pearson, Inc.).

[17] Pentzaropoulos, G.C. (2013). Limits of responsiveness for geographically remote knowledge bases: an operational analysis. *Advanced Modelling and Optimization,* 15 (2), pp. 249-260.