ВЕСТНИК ПЕРМСКОГО УНИВЕРСИТЕТА

2019 **История** Выпуск 3 (46)

УДК 004:9
doi 10.17072/2219-3111-2019-3-159-174

INFORMATION TECHNOLOGY, INFORMATION AND HISTORY

M. Thaller

University at Cologne, Historisch-Kulturwissenschaftliche Informationsverarbeitung, Albertus-Magnus-Platz, D 50931 Köln, Germany
manfred.thaller@uni-koeln.de

One of the truisms of our decade is, that we live in the information age. Indeed information technology influences all disciplines in academia and its nature is central to the discussion of at least a dozen disciplines, from Floridi's philosophy of information to computer science and from quantum physics to library science. That such a large number of fields of research emphasize different concepts is not really surprising; it would be, if they would not. Nevertheless, all of them are supported by one integrated type of information technology, so the underlying concept must be consistent. In the field of history we can show, however, that the classical derivation of practical information technology, derived from Shannon's implicit creation of an equivalence between communication and information, vulnerates the way information is handled in the field. We will discuss how this creates severe limits to the practical application of current information technology and which changes would be needed to support history's problem domain. This is particularly important as experience shows that theoretical discussions of the nature of information, general or within specific knowledge domains, have almost no influence on the development of the actual information technology, unless the theoretical discussions reflect implementation policies.

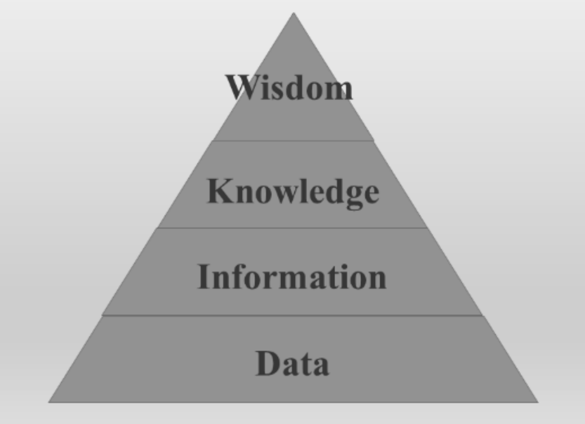
Ключевые слова: Information Technology in History, Source Criticism, Information Theory, Fuzzy Sets, Semantic Computing.

That “information technology” is one of the most central developments of recent times is a well-known truism. That the Humanities in general and Historical research in particular have a tradition of applying information technology within their knowledge domains almost as long as that technology exists, is a truism as well; albeit a less well known one.

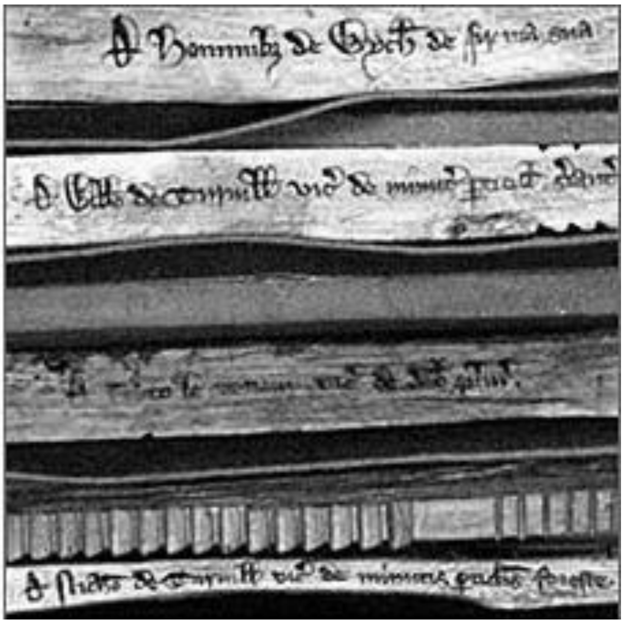
Nevertheless, when you ask what “information” is, the answers provided by different disciplines are very far apart indeed. It is presumably intuitive, that in a discussion of information in quantum theory [*Nielsen, Chuang,* 2000] the explicit concept of information is different from the implicit concept that underlies colloquial discourse about social media. If this is intuitive it should follow easily, that the concepts of information used in different knowledge domains are as different as these knowledge domains themselves. So, the information being processed by the application of information technology to historical sources could be quite different from that underlying information technology for the handling of mundane tasks of daily life. Examining this assumption and some of the consequences is the purpose of this paper.

There *are* definitions of the term, provided by and used within different disciplines. I would like to start with one, which has been used under various names – *Knowledge Pyramid, Ladder of Knowledge, DIKW model* – in various disciplines: information science, philosophy of information, cybernetics and a few others, though rarely in information theory [*Ackoff,* 1989, *Ashenhurst,* 1996, *Rowley,* 2007, *Frické,* 2009, *Saab,* 2011, *Baskarada,* 2013, *Jifa,* 2014, *Duan,* 2017]. And usually *not*, or not much so, in computer science and information technology. The latter two are frequently quite satisfied with being able to represent information in data structures and process them by algorithms, without having to define what it actually is. The model I start with is a compromise between the numerous varieties which have been proposed by different researchers, most closely following [*Favre-Bull,* 2001].

The definitions of these layers will follow soon, let me start with an intuitive example first. Let’s look at some medieval tally sticks. These were short wooden sticks, which were given as acknowledgment of some economic relationship, described in writing on the stick. Notches at the top ensured on the one hand, that the form of the stick was unique, at the other the notches in many cases representing also the amount of what was owed or simply counted. The stick was than split lengthwise and the two parts were kept by the two parties of the transaction. Having an irregular shape, the individual parts could not be tampered with. Graphic 2 shows such a tally stick, holding *information.* The notches on top also represent the amount of what is described in the text.

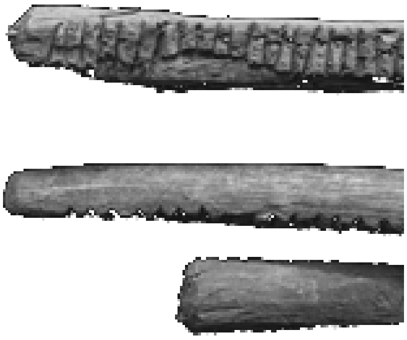


Graphic 1: Knowledge Pyramid



Graphic 2: Information recorded on tally sticks

Graphic 3, on the other hand holds only *data*. Whether the writing was never there, as the two partners of the transaction were illiterate, or whether it has eroded: We see notches, which counted something – or something else. *Which* something, well …

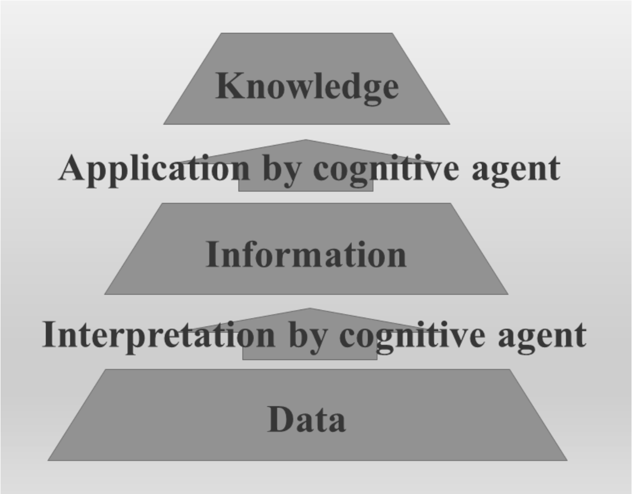


Graphic 3: Data recorded on tally sticks

More formally: *Data* are marks in some representational system, which can be stored. *Information* results, when these marks are put into some context. So "22°" are data. "The temperature of this room is 22°" is information. *Knowledge* arises, when this information encounters the ability to draw advice for action from it. "I do not feel overly warm just because I had to run to get here in time. I really should get out of my jacket."

Important later: Most researchers – notable exception: Luciano Floridi [*Floridi*, 2011, 182 ff.] – do not think, that truth has to do anything with all this. Assume your knowledge of the world includes the "truth" that vampires exist. The data represented by a color change of the garlic in your garden convey the information that it is ready for harvest. Within *your* view of the world, it is valuable knowledge to deduct from that information the plan to use some of it to surround your windows.

Ignoring the rather elusive wisdom layer: How can we represent the relationship between the remaining conceptual layers more stringently? The usual approach is represented in graphic 4.

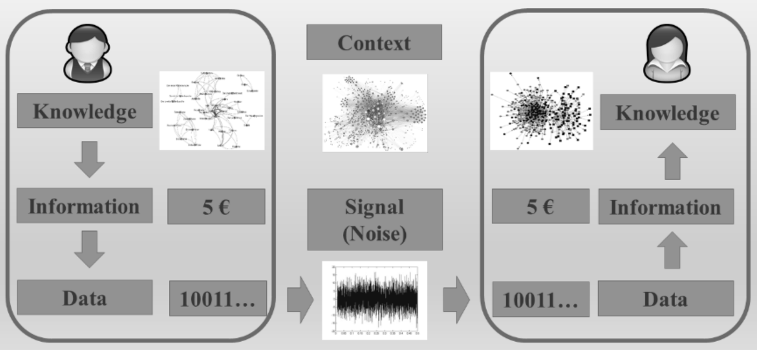

Knowledge
Application by cognitive agent
Information
Interpretation by cognitive agent
Data

Graphic 4: Information among agents

Cognitive agents – you, me, a component of a "smart" piece of software – can perform their activity, as they have other knowledge in the background: That a number you see on a thermometer represents temperature and how to translate that into action requires information or knowledge beyond the number indicated by the device. Both processes have in common, therefore, that they put a specific chunk of data or information into a *context*. There is a very big difference between the two levels, however.

To convert the data "22°" to the information "the temperature of this room is 22°", requires contextual information that this is the way to read a thermometer. This is common to *all* cognitive agents which operate successfully in that environment. The contextual knowledge, that this temperature should trigger the action of getting rid of your jacket, is restricted to one specific cognitive agent, you. So, the context required to convert data into information is shared between a larger number of agents, the context required to convert information into knowledge is private to a specific agent.

If we try to use this model to describe what is happening in communication, we get the following graphic, inspired by [*Favre-Bull*, 2001, p 87, ill. 47]. The icons of graphs represent the contexts. With this in mind, we can describe the situation of two people discussing a possible purchase as illustrated by graphic 5.



Graphic 5: Discussing the price of a purchase

This graphic is to be read as follows: *The person in the top left corner has been asked what the price of an item is, which the person on the top right-hand corner wants to purchase. Based on his knowledge about the price he fixed before or the going rates for such merchandise, he selects an amount. This is converted into information, combining an amount and a currency, and in the next step into data which can be transmitted. Whether the signal used for that purpose is a set of sound waves at a market stall or a string of bits transmitting from an internet shop via the WWW is irrelevant for our purpose. In both cases, the purchaser at the right-hand side will receive that signal, hopefully undistorted by the acoustic noise at the market or the static electricity around the connecting lines. The signal reconstituted as data gets transformed into information as amount and currency and allows the purchaser to act upon the offer, contemplating it in terms of the desirability of the object and her overall budget.*

That all of this works requires that the background knowledge of the two participants in the conversation and the common context of sociocultural conventions between them overlaps sufficiently, that they have the feeling to "understand" the other party. Whether the seller believes to ask an outrageous price, while the purchasing tourist understands she's been made a gracious offer, does not really hinder the action. But incompatible contexts *may* destroy the communication: If the tourist believes she must haggle while the seller offers a fixed price the two parties will not get together.

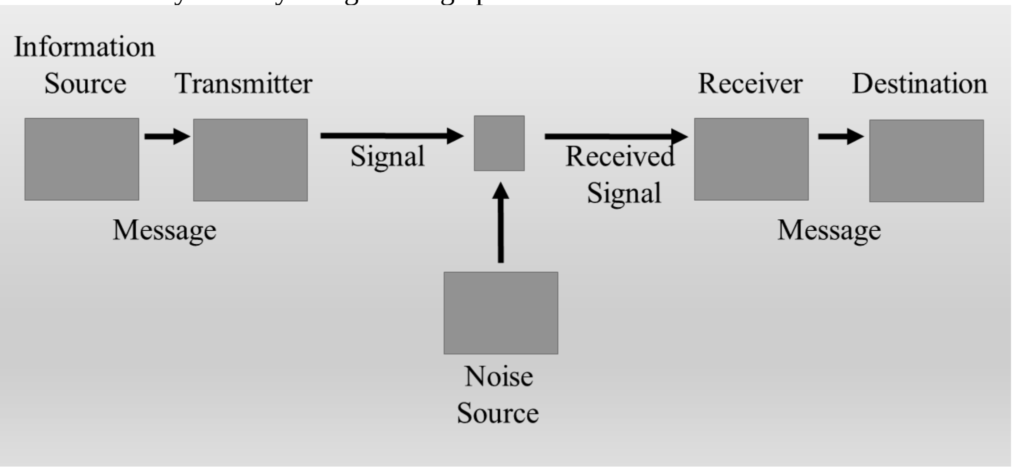
So, there are two levels at which communication may break down: The transmission of the signal amidst noise at the bottom of the diagram and the compatibility of the knowledge context – semantic context, for short – on top.

One might say that the latter is the most crucial problem when applying this model to information as contained in historical sources. If you replace in graphic 5 the present day icons of seller and purchaser by photographs of statues of Pericles and Thucydides and the string "5 €" by the string "πόλεμος" the diagram is still completely appropriate to describe what Thucydides understood, when listening to Pericles explaining his policy, as both shared the same sociocultural context. If you replace the seller by Pericles and assume that the lady's icon at the right-hand side represents a modern historian, you immediately see that this will not work – a modern historian simply does *not* share the sociocultural context with Pericles. Which is an incomparably more serious problem, than the question how close the text of Thucydides we have is to what he actually wrote, leave alone what words Pericles really had chosen.

Nevertheless, information technology in its broadest sense until quite recently had the tendency to focus on how to overcome the signal noise, not so much the semantic noise of a process of communication. This goes back to the seminal paper by Claude Elwood Shannon published in two parts in 1948, entitled *A Mathematical Theory of Communication* [Shannon 1948]. Shannon was very clear on what he thought could be done and what could not be done. The second paragraph of his paper starts:

"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.*" ([*Shannon*, 1948, p. 379] "Meaning" italicized by Shannon; last sentence by me.)

This as an introduction to the basic diagram of the model of communication as reproduced in graphic 6, the traces of which you easily recognize in graphic 5.



Graphic 6: Shannon's model of communication

Shannon’s paper was most fundamental to modern information technology; but it was not exactly easy to read. It was republished as a book already one year later in 1949, entitled *The Mathematical Theory of Communication* this time [*Shannon,* 1949]. To ease understanding it was introduced by an introduction by Warren Weaver, as mathematician highly qualified to understand the original argument and personally highly qualified to write transparent prose, so his text probably influenced the public perception of Shannon’s much more than the original text. He lists three levels of communication problems, which easily relate to our concepts of data, information and knowledge:

“Level A. How accurately can the symbols of communication be transmitted? (The technical problem.)

Level B. How precisely do the transmitted symbols convey the desired meaning? (The semantic problem.)

Level C. How effectively does the received meaning affect conduct in the desired way? (The effectiveness problem.)” [*Weaver,* 1949, p. 24]

Two pages later, Weaver even writes:

“So stated, one would be inclined to think that Level A is a relatively superficial one, involving only the engineering details of good design of a communication system; while B and C seem to contain most if not all of the philosophical content of the general problem of communication.” [Weaver 1949, 4] Which basically rephrases Shannon’s statement.

But barely two pages later he inexplicably continues:

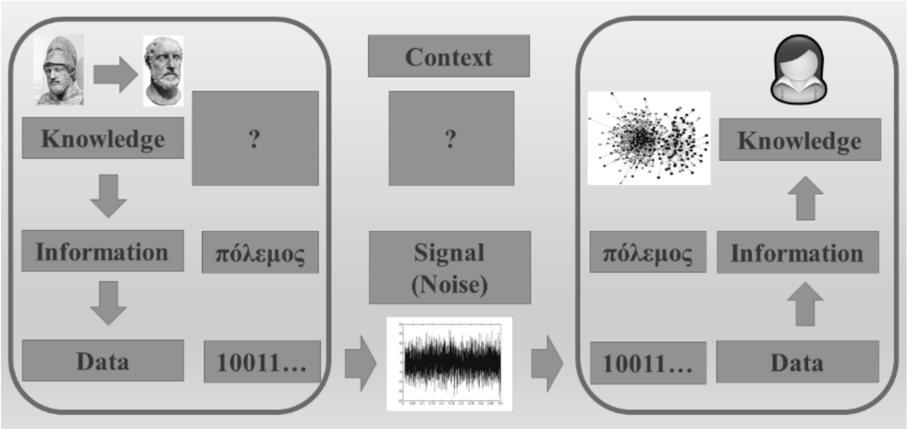
“Part of the significance of the new theory comes from the fact that levels B and C, above, can make use only of those signal accuracies which turn out to be possible when analyzed at Level A. Thus, any limitations discovered in the theory at Level A necessarily apply to levels B and C. But a larger part of the significance comes from the fact that the analysis at Level A discloses that this level overlaps the other levels more than one could possible [sic] naively suspect. Thus, the theory of Level A is, at least to a significant degree, also a theory of levels B and C.” [*Weaver*, 1949, p. 6]

And he starts his conclusion triumphantly:

“It is the purpose of this concluding section to review the situation and see to what extend and in what terms the original section was justified in indicating that the progress made at Level A is capable of contributing to levels B and C, was to indicating [sic] that the interrelation of the three levels is so considerable that one’s final conclusion may be that the separation into the three levels is really artificial and undesirable.” [*Weaver,* 1949, p. 25]

This seems to be at the root of the popular perception of Shannon’s model. If we ignore this misinterpretation of Shannon and emphasize the difference between the various layers of the model we started from, we should next apply the model of figure 5 to the domain of historical research. “Historical research” understood as the process deriving information from sources and integrating the greatest amount of available information into a consistent model of the historical processes that produced these sources.

So let us start our consideration of the information contained in historical sources with graphic 7, modified from graphic 5 to apply that model to the situation a historian finds herself in, when she tries to understand Pericles’ reasons to implement his policy regarding the war with Sparta, as reflected by Thucydides.

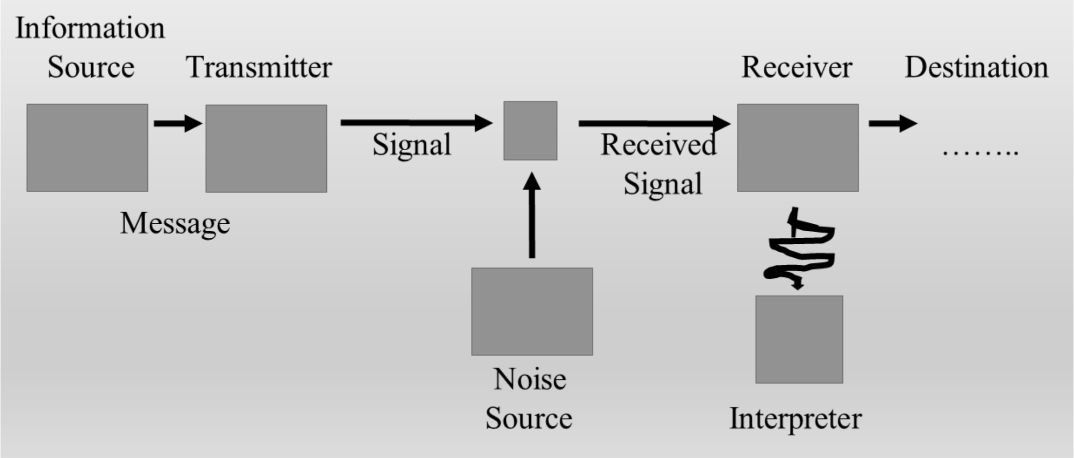


Graphic 7: Reasoning about Pericles' policies

Our friend faces two problems: The first we already mentioned, when discussing information as such. We do not share the silent assumptions of the sociocultural context (question mark in the middle), so we have no access to the context in which Pericles did formulate his policies. But we have ignored another problem so far: The information we have did not originate from Pericles, but from Thucydides, being the result of an earlier communication process. Estimating the distortions produced by such earlier communication processes is the bread and butter of historical research, as far as it is focused on the content and not the literary qualities of historical writing.

Since the 19th century this has at least in the German tradition been clearly fixed in historical methodology. Droysen [Droysen 1937, 38-50] clearly emphasized that traditions (*Traditionen*), like chronicles, which were the result of an intentional effort to leave a specific view of an event, were less valuable as sources than remainders (*Überreste*) which resulted from processes which were not controlled by an intention to leave a specific image for the coming generations.

For us this implies, that a historian is *not* interested in the message the author of a source *wanted* to transmit, but rather in such insights about the situation, which a source provides as *independent* of the *intentions* of the author. Historians do not receive messages: they use them to reconstruct lost contexts – always tentatively. Which graphic 8 tries to visualize.



Graphic 8: Transmission model for historical sources

Let me close this section with a high-level view of the requirements of historical information systems, which I consider to be derived from our musings on the nature of information above.

An information technology appropriate for historical sources:

1. Represents the artifacts as free from any interpretation as possible in the technical system,
2. embeds them, however, in a network of interpretations of what they imply,
3. provides tools which help to remove contradictions between such interpretations,
4. accepts, however, that such contradictions may prove to resist resolution
5. as well as that all interpretations always represent tendencies, no certainties.

While to the best of my knowledge the above attempt to intertwine musings about historical sources with considerations of the nature of information as handled by technical systems is a rather rare exercise, musings about the nature of information in such systems are anything but rare. As mentioned at the very beginning, the knowledge pyramid has left traces in information science, cybernetics, the philosophy of information and a few others. But with all respect due to the excellent work done by all these disciplines, it is hard to see, how information technology would have developed differently from the way it did, if these disciplines would never have existed. Extremely rare exceptions – Norbert Wiener, e.g. – prove the rule. The main systematic exception is cognitive science, which contributes its own share of theoretical reflections on the nature of information, frequently interconnects with the implementation of technical solutions however, and *therefore* actually influences technical development. If thinking about the nature of information in historical sources does not influence the development of information technology this means that information technology serves historical research less than it should. If historical information can only be handled by systems which cannot handle contradictory information, such systems are ultimately inappropriate for this usage.

In software engineering, we are familiar with the term “technology stack” or “software stack” describing the selections of technologies made at various levels to implement a system. A frequent example defines a stack as the choice of an operating system, a web server, a database system and a programming language. LAMP (Linux, Apache, MySQL, {Perl, PHP or Python}) is the best-known example. The term “stack” leads to the association, that once the bottom level has been chosen, the choice of the upper levels is severely restricted. The existence of the WAMP (Windows, Apache, MySQL, {Perl, PHP or Python}) stack shows, however, that the different levels *are* related, but not strictly hierarchical. I propose *conceptual stack* as a new term for the combination of general concepts which go – implicitly or explicitly – into the design of information systems. A conceptual stack in that sense is more abstract than the software stack, but sufficiently concrete to determine specific properties and capabilities of all information systems build upon that stack. As in the technical case, we assume these concepts to depend on each other, but not in a strictly hierarchical sense.

In contemporary information systems, I identify at least five conceptual decisions, which restrict their usefulness for the handling of information contained in historical sources as discussed above.

These are:

(1) The interpretation of the signals used for communication with the concept of granular units of information. What in my understanding is *Weaver’s distortion* of Shannon’s concepts.

(2) The believe that, as bits can be used to conveniently implement Boolean logic, computer systems necessarily have to be based on binary logic. What in my understanding is the *binary fallacy*.

(3) The assumption that the language of historical documents can best be approached by analyzing their syntax. What in my understanding is *Chomsky’s dead end*.

(4) The approach to embed interpretations of an object into their representation. What in my understanding is the *markup fallacy*.

(5) The principle, that variables in a programming language are conceptually independent of each other, as long as they are not explicitly connected into a structure or object. What for reasons not immediately apparent is the *Gorilla syndrome* in my understanding. As this is the most technical of the five, no details are given below.

I will try to describe them briefly and give hints how implementations of solutions could look like. As the readers of this journal will mainly be interested in the historical parts of this paper, I provide only rather sketchy descriptions of the technical solutions, avoiding as many technical details as possible.

Weaver’s Distortion

My accusation against Weaver that in the attempt to make Shannon’s model easier to understand he mixed up different conceptual levels, can probably be made a bit more transparent by an analogy. In the physical world we are perfectly aware, that there exist two closely connected but, in many ways, independent sub-worlds: A Newtonian one and a world that is ruled by Quantum physics. They are closely connected; nevertheless, the confusing habits of quarks do not prevent the Earth to circle the sun in an encouragingly reliable way – even if gravitation, responsible for the reliability, can probably be understood only on the sub-nuclear level. My proposal is, that a similar separation can be used to understand the relationship between the world of data, turning into information and knowledge in the context of other data, and the signals constituting those data. On the possibility to use more than one theory of information in parallel see [Sommaruga 2009].

The computational legacy of Weaver’s distortion is, therefore, inherited by the programming paradigms we use today. So, operations which shall interpret *information* are handled by data types, which are exactly that: *data*.

One of the reasons for this is, that what you do with numbers can be understood and validated by the formal apparatus provided by analysis in the mathematical sense and particularly by numerical analysis among its branches. For strings similarly stringent formalisms have been developed.

But anything related to meaning is more slippery. Though there have been attempts to change that: Keith Devlin specifically proposed a mathematical theory which addresses that problem:

“… whereas in this essay I am taking information itself as the basic entity under consideration. More precisely, I am seeking a specific conceptualization of ‘information’ as a theoretical ‘commodity’ that we can work with, analogous to (say) the *numbers* that the number-theorist works with or the *points*, *lines* and *planes* the geometer works with." [*Devlin,* 1991, p. 17] (All italics are Devlin's.)

Staying at the lowest illustrative level [cf. *Devlin,* 2009] we can describe his approach by the notion of an "infon" as the atomic unit of an information system. An infon is defined as

<< P, a_1, …, a_n, l, t, i >>

The parameters are defined as:

P – an n-ary relation

a_1, …, a_n – objects between which P holds

l – a spatial location

t – a temporal location

i – a truth value

As an example, Devlin [*Devlin*, 1991, p. 24] gives:

<< marries, Bob, Carol, *l, t,* 1 >>

to describe the information, that Bob marries Carol at a location *l* and a date *t*. As this is true, the final parameter is 1. If it would be wrong, it would be 0.

Considering this as a basic notion of information has many attractions. It acknowledges, that information grows out of data in context and it reflects that knowledge does not have to be true. Zeus and Hera *are* married on Olympus, even if they do not exist and the spatial location of their Olympus requires an interesting extension of the concept of space.

Research proposal in software technology 1:

Implement infons for seamless usage in mainstream programming languages.

But our argument for the basing of information systems on other building blocks than the current data types does not end here. We have derived this requirement from our initial consideration that information should be understood in the sense of one or the other version of the knowledge pyramid, not on the level of signals.

For simplicity's sake, we have so far avoided the discussions of the shortcomings of the knowledge pyramid itself, which has led to various criticisms against it and occasional calls for its abolishment. Let's look again at the starting example.

Information results, when these marks are put into some context. So "22°" are data. "The temperature of this room is 22°" is information.

However: What about the number "22"? Before it turns up on a thermometer, it might measure the length of a room in feet, the weight of truck in tons, the distance of two towns … So: "22" is data; by being contextualized as "22°" it becomes information. And what about the bit string "0000001000000010"? It could be the "device control character 2" from the ACII table, the first half of the Unicode code for the universal quantifier symbol, … So: "0000001000000010" is data; by being contextualized as "the value of an integer variable" it becomes information. And so on.

We can argue that this contradicts the knowledge pyramid model; or that it emphasizes its validity as a confirmation of the overwhelming importance of context, even if the transitions between interpretative levels are more complex than the simple 3 level model – data, information, knowledge – indicates.

I recommend that we solve this by an approach for the understanding of the concept of information, which unfortunately has left only very few traces [e.g. *Kettinger,* 2010] in the current discussion: Langefors' "infological equation" [*Langefors,* 1973]. According to him, the information communicated by a set of data, is understood to be a function i() of the available data D, the existing knowledge structure S and the time interval t, which is allowed for the communication, given by the formula

$I = i(D,S,t)$

The interesting thing about Langefors' equation is, that it introduces the time a communication – or the process of generating information out of data – takes. If information is derived from data in a continuous process, we must assume, that the longer that process may take, the more information we may extract – "more information" easily conceptualized as "information in a more complex context". Formulated differently: If this is a *process* represented by a function, rather than a timeless transition, there is not a discrete, but a continuous relationship between data and information. That is, what has been data at one stage, is information at another. We can therefore write

$I_2 = i\ (I_1, S_2, t)$

to represent the notion, that the information available at time “2” is a function of the information available at time “1” and the knowledge available at time “2”, depending on the length of time we have available for the execution of that function. As I have discussed this in detail somewhere else [*Thaller,* 2009, 345ff.] I will cut the argument short and just mention, that from this stage we can develop the argument further arriving at a model, which assumes that no such thing as static information exists; “representing it” just captures a snapshot of a continuously running algorithm.

Research proposal in software technology 2:

Represent information as a set of conceptually permanently running algorithms, the state of which can be frozen and stored.

The Binary Fallacy

The practical training of computer scientists and software technologists starts by emphasizing binary concepts. This creates a distorted look upon what is actually happening in software systems.

If you look at the software fragment

```
int parameter;
…
if (parameter) doSomething();
```

most programmers would spontaneously translate the condition as “if parameter is true or one, execute the function. If it is false or zero, don’t.” Only if pressed, they would give the correct reading “if parameter has any other value but zero – that is: $-2^{15}+1$ till -1, 1 till $2^{15}-1$ – execute the function.

That is, a Boolean interpretation of the code fragment actually underuses what it represents. Therefore, there is *no* technical reason why computations should be restricted to binary logic, *nor* is there any compelling reason, why a “number” must be a zero-dimensional point on the continuum.

Both of these preliminary observations should be kept in mind, when we think about the inherent fuzziness of information derived from our metaphor of the historian as observer of signals once exchanged between participants in a context which has been lost.

There are a number of phenomena which this inherent fuzziness encompasses. Without claiming completeness: (1) Fuzziness in a narrower sense, i.e., the impossibility to give a crisp truth value for a statement. (2) An inherent imprecision of a semantic concept, as in “old people”. (3) An item which conceptually is a scalar but goes beyond our current data types. E.g. a price for a commodity, for which we do not have a precise value, but a minimum and a maximum, plus possibly hints at the distribution of the data points between these.

For all three of the problems mentioned partial solutions exist. Usually so high up the technology stack however, that they are applicable only under very special circumstances. To support their general application, they would have to be provided at the level of higher programming languages.

(3) “An item which conceptually is a scalar but goes beyond our current datatypes. E.g. a price for a commodity, for which we do not have a precise value, but a minimum and a maximum, plus possibly hints at the distribution of the data points between these.”

For this problem, Julong Deng’s Grey System Theory, or rather the systematic treatment of the building blocks for such systems as described by Sifeng Liu and Yi Lin [*Liu,* 2006, 2011], seems to provide almost a blueprint for implementation.

Research proposal in software technology 3:

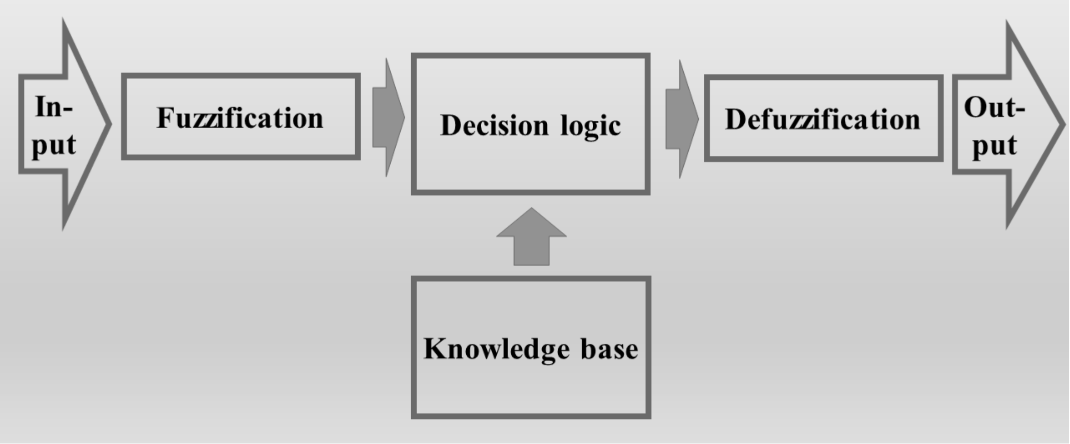
Implement grey numbers, or a derivation from them, and integrate them seamlessly into mainstream programming languages.

(2) “An inherent imprecision of a semantic concept, as in ‘old people’.”

It is obvious, that this problem closely relates to Zadeh’s [1965, 1975, 1978, 1999] seminal work on Fuzzy Sets and Systems, and the later concept of “Computing with Words” based on linguistic variables, which has found widespread applications in many branches of computation. In almost all of them, actually, except in the Humanities – which Zadeh himself originally expected to be primary fields of application [*Blair,* 1994; *Termini,* 2012].

The smaller generalization required is, that since 1965 Zadeh’s Fuzzy Sets similar concepts have been explored [*Pawlak,* 1982, 1985; *Shafer,* 1976] and an almost endless list of modifications of the basic approaches [*Atanassov,* 1986; *Torra,* 2010; *Herrera,* 2014; *Nanda,* 1992; *Jiang,* 2009] has arisen. Zadeh himself in his later years has tried to combine some of these approaches into a Generalized Theory of Uncertainty [*Zadeh*, 2005] but this is quite restricted in scope. As Barr has noted "One of the problems with fuzzy sets is that the meaning of the term has been left vague (one might say fuzzy)." [*Barr*, 2010, 400]

A more general problem is, that most of the approaches I have encountered, still assume fuzziness to be the exception, rather than the rule, as it has to be, if we accept the model of an observer of signals exchanged in a lost context. The current logic of embedding an approximately reasoned decision into an information system, is illustrated by graphic 9, an attempt to generalize the similar graphics contained in the literature.



Graphic *9*: General logic of "computing with words"

With other words: From an information system, which is crisp, some information is transferred into a fuzzy box, the result of the decision made crisp again for the major parts of the larger embedding system.

Research proposal in software technology 4:

Implement linguistic variables and integrate them seamlessly into mainstream programming languages, as permanently accessible data type in all parts of the flow of execution. Base the implementation on a generalized concept of uncertainty, which broadens the scope of Zadeh's theory of that name.

(1) "Fuzziness in a narrower sense, i.e., the impossibility to give a crisp truth value for a statement."

This is in some ways the most puzzling problem for software technology. At first look it seems to be rather simple, as logics with multiple values of truth, preferably continuous truth functions, are well understood and an ample literature exists. The engineering problem appears, however, when the evaluation of an expression in multivalued logic is the base of a control structure.

In the code fragment

```
if (condition_is_valid) doSomething();
else doSomethingElse();
```

what happens, if "condition_is_valid" has a truth value of 0.75?

To the best of my (admittedly incomplete) knowledge a very early proposal for the inclusion of fuzzy logic into a programming language – Adamo's LPL [*Adamo*, 1980] – is the only one, where for such cases a combination of the execution of both branches is contemplated in detail.

Research proposal in software technology 5:

Design genuinely fuzzy control structures and integrate them seamlessly into mainstream programming languages.

Chomsky's Dead End

Linguistics and programming languages have shared an intimate relationship for a long time. This has led to a focus of linguistic work on syntax which is at the least not very productive for historical work, if not outright counterproductive.

"*Igitur Carolus Magnus a Leone III. Pontifice Romae anno 800 coronatus*, …" ("So Charlemagne was crowned in the year 800 in Rome by Pope Leo III, …"; Robert Bellarmine (1542-1621), *De Translatione Imperii Romani*, liber secundus, caput primum)

Charlemagne, Leo III and cardinal Bellarmine all would have been able to "understand" this sentence, as all of them were quite familiar with Latin grammar. But what would they really have understood? The chief of a network of post-tribal Germanic loyalties and the son of a modest family in Southern Italy, where Byzantium was still looming large, *might* have had at least similar notions of what the title “Imperator” implied. The idea which the highly educated 16th/17th century cardinal connected with the title would certainly have been incomprehensible for both – and vice versa.

Whether the fixation on syntax is good for linguistics, is a question linguists will have to answer. Why a semantic understanding requires an understanding of the syntax of a message, at least I have never understood. There are linguists which doubt it – formulated by Roy Harris as “Do we always know what we mean?” [Harris 1998, 14] Indeed, do we? Are we aware of all the assumptions we make in formulating a sentence and all the implications our linguistic choices make for the listener? If so, how can we hope ever to understand an utterance where the last native speaker has died a few hundred years ago?

Of the five problems I mention, this is probably the one, where it is hardest to plausibly describe in brief a direction for a technical – algorithmic – solution.

Conceptual alternatives exist: The notion of Lakoff and Johnson, that understanding is based upon metaphors [*Lakoff*, 1980] is immediately intuitive for a historian who has tried to look through the meaning hidden behind historical texts. And the notion, that the possibility to construct associations between different concepts, to *blend* concepts [*Fauconnier*, 2003] is actually a feature distinguishing the human mind in a much more fundamental way, than the I-Language and the Universal Grammar [*Isac*, 2008], is highly convincing for the same historian.

But, as I said, while for most of the other proposals to solve a concrete technical problem in this paper, a sensible starting point and the first few steps of the solution are clear, how to implement metaphors and conceptual blending is much harder to see. What might be very useful as the starting point would be semantic graphs, which allow the handling of seemingly contradictory relationships between nodes. Blending of two concepts means, that an edge connects two nodes, where from the point of view of any of the two concepts, no edge should exist. This would require a class of graphs were both nodes and edges are labeled and there exist bundles of edges, where the acknowledgment of one implies the rejection of one or more of the others. These would also be useful for the implementation of the type of graphs discussed in the next section.

Research proposal in software technology 6:

Provide tools for the easy handling of such graphs in mainstream programming languages.

The Markup Fallacy

Markup languages are not a very central subject of computer science. For many humanists, however, they seem to be the quintessence of the so-called Digital Humanities. In my opinion, they current usage of markup in the handling of historical documents has two methodological weaknesses.

The first of these is rather straightforward: Embedding markup into a text goes directly against the principle mentioned earlier, that a software system handling historical information “Represents the artifacts as free from any interpretation as possible in the technical system …”. On the surface that is violated by some principles, which the TEI has propagated strongly in its early days, when markup should signal meaning, not display features, resulted in the idea that e.g. italics should result in an <emph> </emph> tag. As “meaning” is an interpretation, not a representation of the source, this violates my methodological understanding of historical research.

But there is a much more fundamental problem with this sort of markup, when one looks at it from the point of view of processing historical data in information systems. “Markup” according to the current paradigm, applies to text; adding explanatory or analytic comments to an image, a 3D reconstruction or any other non-textual material is considered an “annotation”. (Though annotations have recently also appeared related to text.) I can see no epistemological reason whatsoever, why texts and other forms of source representations are handled differently.

In principle it is quite possible to define standoff annotations which provide a homogeneous solution for one-dimensional (textual), two-dimensional (images) … n-dimensional data. Indeed, in the context of long-term preservation, we have proven that this is technically viable in one of my earlier projects [Thaller 2009].

However, while Desmond Schmidt [Schmidt 2009] has proposed a solution for preparing standoff markup for a text, in a way which allows editing of the text independent of the markup, I am not aware of any solution, which would allow this for a data object of higher dimensionality.

Research proposal in software technology 7:

Develop a representation of “information objects”, where a data object of arbitrary dimensionality can be combined with interpretative layers in such a way, that the data object can be changed without damaging these layers.

There is a small *caveat* to be added to the above. All of these considerations relate to the situation, where a source is converted by a 1 : 1 operation into a technical representation, be it a human transcription or the scanning operation of an image. Despite the emphasis on leaving a source as undistorted as possible, there is of course the need to handle data objects which represent attempts to create a common representation of more than one such object, e.g. the reconstruction of the commonalities between the witnesses of a an abstract text surviving in different manuscripts. The “leave the source unchanged” principle should apply here as well. For such nonlinear texts the equation “one source is represented as a string, i.e., an array of characters” obviously does not hold. I have myself proposed a model for representing texts not as arrays, but as graphs [*Thaller,* 1993]. Similar situations may become important in the future in data objects of other dimensionality, when the explosive spread of scanning techniques emphasizes more strongly the need to represent relationships between families of images or other objects of higher dimensionality.

Conclusion

Information technology, as it is represented by the current technical mainstream, is built solidly upon a number of assumptions which go back to the theoretical foundations of signal theory and the various main areas of knowledge domains which have been important in the development of such technologies. Information technology assumes, e.g., that information: is generated by processes, where the meaning of the symbols used can always be enquired about from the sender; can be mapped unto crisp categories within a small and in any case finite number of steps; fits into canonical data structures like graphs as supported by current implementations of relevant libraries of functions. We have claimed, that information as it is derived from historical sources, vulnerates all of these assumptions. The meaning of the symbols of such information can not be verified with the sender, they will always carry a degree of uncertainty; the categories into which the can be mapped are inherently fuzzy and such mappings may require permanent reiterations; resulting in processes, which need extensions and generalizations of existing data structures and the software supporting them. A number of concrete developments to achieve optimized support for the handling of historical information under such conditions have been identified.

References

Ackoff, R.L. (1989), “From Data to Wisdom”, *Journal of Applied Systems Analysis*, vol. 15, pp. 3-9.

Adamo, J.M. (1980), “L.P.L. A fuzzy Programming Language: 1 Syntactic Aspects,” *Fuzzy Sets and Systems*, vol. 3, pp. 151-179.

Adamo, J.M. (1980), “L.P.L. A fuzzy Programming Language: 2 Semantic Aspects,” *Fuzzy Sets and Systems*, vol. 3, pp. 261-289.

Ashenhurst, R.L (1996), “Ontological Aspects of Information Modeling”, *Minds and Machines*, vol. 6, pp. 287-394.

Atanassov, K.T. (1986), “Intuitionistic Fuzzy Sets”, *Fuzzy Sets and Systems*, vol. 20, pp. 87-96.

Barr, M. & C. Wells (2010), *Category Theory for Computing Science*, Montréal, Canada.

Baskarada, S. & A. Koronios (2013), “Data, Information, Knowledge, Wisdom (DIKW): A Semiotic Theoretical and Empirical Exploration of the Hierarchy and its Quality Dimension”, *Australasian Journal of Information Systems*, vol. 18, pp. 5-24.

Blair, B. (1994), “Interview with Lotfi Zadeh”, *Azerbaijan International*, vol. 2, Winter, pp. 46-47, 50.

Devlin, K. (1991), *Logic and Information*, Cambridge, UK.

Devlin, K. (2009), “Modeling Real Reasoning”, in: Sommaruga, G. (ed.): *Formal Theories of Information*, (= Lecture Notes in Computer Science 5363), Berlin-Heidelberg, Germany, pp. 234-252.

Droysen, J.G. (1937), *Historik. Vorlesungen über Enzyklopädie und Methodologie der Geschichte,* ed. by Rudolf Hübner, München, Deutschland.

Duan, Y. et al. (2017), “Specifying Architecture of Knowledge Graph with Data Graph, Information Graph, Knowledge Graph and Wisdom Graph”, presented at SERA 2017, available at: doi.org/10.1109/SERA.2017.7965747 (accessed 10.07.2019).

Fauconnier, G. & M. Turner (2003), *The Way We Think. Conceptual Blending and the Mind’s Hidden Complexities*, New York, USA.

Favre-Bull, B. (2001), *Information und Zusammenhang. Informationsfluß in Prozessen der Wahrnehmung, des Denkens und der Kommunikation*, Heidelberg, Deutschland.

Floridi, L. (2011), *The Philosophy of Information*, Oxford, UK.

Frické, M. (2009), “The Knowledge Pyramid: A Critique of the DIKW Hierarchy”, *Journal of Information Science*, vol. 35, pp. 131-142.

Harris, R. (1998), *Introduction to Integrational Linguistics*, Oxford, UK.

Herrera, F. et al. (eds.) (2014), “Special Issue on Hesitant Fuzzy Sets”, *International Journal of Intelligent Systems*, vol. 29, pp. 493-595.

Isac, D. & C. Reiss (2008) *I-Language*, Oxford University Press, Oxford, UK, 392 p.

Jiang, Y. et al. (2009), “Reasoning with Expressive Fuzzy Rough Description Logics”, *Fuzzy Sets and Systems*, vol. 160, pp. 3403-3424.

Jifa, G. & Z. Lingling (2014), “Data, DIKW, Big Data and Data Science”, *Procedia Computer Science*, vol. 31, pp. 814-821.

Kettinger, W.J. & Y. Li (2010), “The infological equation extended: towards conceptual clarity in the relationship between data, information and knowledge”, *European Journal of Information Systems*, vol. 19, pp. 409-421.

Lakoff, G. & M. Johnson (1980), *Metaphors We Live By*, Chicago, USA, with a substantial afterword reprinted 2003.

Liu, S. & Y. Lin (2006), *Grey Information. Theory and Practical Applications*, London, UK.

Liu, S. & Y. Lin (2011), *Grey Systems. Theory and Practical Applications*, London, UK.

Langefors, B. (1973), *Theoretical Analysis of Information Systems*, Göteborg, Germany.

Nanda, S. & S. Majumdar (1992), “Fuzzy Rough Sets”, *Fuzzy Sets and Systems*, vol. 45, pp. 157-160.

Nielsen, M.A. & I.L. Chuang (2000), *Quantum Computation and Quantum Information*, Cambridge, UK.

Pawlak, Z. (1982), “Rough Sets”, *International Journal of Parallel Programming*, vol. 11, pp. 341-356.

Pawlak, Z. (1985), “Rough Sets and Fuzzy Sets”, *Fuzzy Sets and Systems*, vol. 17, pp. 99-102.

Rowley, J. (2007), “The Wisdom Hierarchy: Representations of the DIKW Hierarchy”, *Journal of Information Science*, vol. 33, pp. 163-180.

Saab, D.J. & U.V. Riss (2011), “Information as Ontologization”, *Journal of the American Society for Information Science and Technology*, vol. 62, pp. 2236-2246.

Schmidt, D. & R. Colomb (2009), “A Data Structure for Representing Multi-Version Texts Online”, *International Journal of Human-Computer Studies*, vol. 67, pp. 497-514.

Shafer, G. (1976), *A Mathematical Theory of Evidence*, Princeton University Press, Princeton, USA, 314 p.

Shannon, C.E. (1948), “A Mathematical Theory of Communication”, *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656.

Sommaruga, G. (2009), “One or Many Concepts of Information?”, in Sommaruga, G. (ed.), *Formal Theories of Information*, (= Lecture Notes in Computer Science 5363), Berlin-Heidelberg, Deutschland, pp. 253-267.

Termini, T. (2012), “On some ‘Family Resemblances’ of Fuzzy Set Theory and Human Sciences”, in: Seising, R. & V. Sanz (eds.), *Soft Computing in Humanities and Social Sciences* (= Studies in Fuzziness and Soft Computing 273), Berlin-Heidelberg, Deutschland, pp. 39-54.

Thaller, M. (1993): “Historical Information Science: Is there such a Thing? New Comments on an Old Idea.”, in Orlandi, T., *Seminario discipline umanistiche e informatica. Il problema dell' integrazione*, Rome, Italy, pp. 51-86. Reprinted under the same title in: Historical Social Research, Suppl. 29 (2017), pp. 260-286, available at: doi.org/10.12759/hsr.suppl.29.2017.260-286 (accessed 10.07.2019).

Thaller, M. (2017), “The Cologne Information Model: Representing Information Persistently”, in Thaller, M. (ed.), *The eXtensible Characterisation Languages – XCL*, Hamburg, Deutschland, pp. 223-39. Reprinted under the same title in: Historical Social Research Supplement 29, pp. 344-356, available at: doi.org/10.12759/hsr.suppl.29.2017. 344-356 (accessed 10.07.2019).

Torra, V. (2010), “Hesitant Fuzzy Sets”, *International Journal of Intelligent Systems*, vol. 25, pp. 529-539.

Weaver, W. (1949), “Introductory Note on the General Setting of the Analytical Communication Studies”, in Shannon, C.E. & W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana and Chicago, USA.

Zadeh, L.A. (1965), “Fuzzy Sets”, *Information and Control*, 8, pp. 338-353.

Zadeh, L.A. (1975), “The Concept of a Linguistic Variable and its Application to Approximate Reasoning”, I – III, *Information Sciences*, vol. 8, pp. 199-249, 301-357 and vol. 9, pp. 43-80.
Zadeh, L.A. (1978), “Fuzzy Sets as a Basis for a Theory of Possibility”, *Fuzzy Sets and Systems,* vol.1, pp. 3-28.
Zadeh, L.A. & J. Kacprzyk (eds.) (1999), *Computing with Words in Information / Intelligent Systems* I and II (= *Studies in Fuzziness and Soft Computing*, vols. 33 and 34).
Zadeh, L.A. (2005), “Toward a Generalized Theory of Uncertainty (GTU) – an outline”, *Information Sciences,* 172, pp. 1-40.

Дата поступления рукописи в редакцию 10.07.2019

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ, ИНФОРМАЦИЯ И ИСТОРИЯ

М. Таллер

Университет Кельна, Historisch-Kulturwissenschaftliche Informationsverarbeitung, Albertus-Magnus-Platz, D, 50931 Кельн, Германия
manfred.thaller@uni-koeln.de

Хотя существует всеобщее согласие о том, что мы живем в «веке информации», отсутствует какое-либо общее определение центрального термина «информация». В разных дисциплинах используются разные концепции, которые обычно отражают эпистемологические допущения этих дисциплин. Удивительно, что внутри информационных *технологий* дискуссии о правильном определении и характере информации довольно редки.

Одна концепция информации использовалась под различными названиями – Пирамида знаний, Лестница знаний, *модель DIKW* (data, information, knowledge, wisdom – данные, информация, знания, мудрость) *(Knowledge Pyramid, Ladder of Knowledge)* – в различных дисциплинах. Короче говоря, предполагается, что информация должна пониматься как часть иерархии понятий, используемых для усвоения и передачи того, что мы наблюдаем. Эта иерархия начинается с *данных*: например, число 22. Когнитивный агент – человек или система программного обеспечения - может преобразовать это в *информацию*, интерпретируя ее в общем контексте. Температура 22 ° – это нечто совершенно отличающееся от возраста 22 лет. Тот же самый или другой когнитивный агент может превратить это в *знание*, позволяющее надлежащим образом реагировать на такую информацию, которая зависит от *частного контекста*, характерного только для одного агента: достаточно ли информации о том, что комната имеет температуру 22°, чтобы снять пиджак, полностью зависит от индивидуальных особенностей.

Тем не менее технические решения проблем с информацией в подавляющем большинстве основаны на надлежащем способе обработки *данных*. Это немного удивительно, поскольку признанный отец современных информационных технологий Клод Э. Шеннон в своей основополагающей работе совершенно ясно говорит [*Shannon*, 1948, p. 379], что его трактовка теории общения ограничивается только аспектом успешного обмена сигналами – данными – между техническими системами, и процесс коммуникации включает в себя различные семантические и концептуальные уровни, однако «эти семантические аспекты коммуникации не имеют отношения к технической проблеме» [*Shannon,* 1948, p. 379].

Крайне прискорбно, что во время популяризации инженерных решений Шеннонса для коммуникационных технологий, которые, несомненно, являются одними из важнейших предпосылок нашего века информации, эта модерация была отменена, и сложилось впечатление, что прогресс, достигнутый в обработке данных, мог бы быть непосредственно и немедленно отражен прогрессом на уровнях информации и знания.

На теоретическом уровне это означает, что одна из самых главных метафор, определяющих наше понимание информационных технологий, не описывает должным образом то, что делают историки, когда обрабатывают информацию. В повседневных процессах общения мы используем наше понимание современного социального и концептуального мира, собственного современного контекста, чтобы расшифровать сообщения от других людей, разделяющих этот контекст. Для историков современный контекст прошлых веков был утрачен априори. Они интерпретируют данные сообщений, которыми обменивались ранее, чтобы реконструировать контексты предыдущих периодов и обществ.

Можем ли мы применить технологию, скрытые предположения которой нарушают концептуальную основу, на которой мы работаем? Мы, конечно, можем, когда используем это для выполнения задачи, не относящейся к историческим исследованиям: для написания статей, например. Однако анализ исторических источников, по моему мнению, являющийся по-прежнему более важным для исторических исследований, чем плавное представление результатов, пострадает. Однако можно заметить, что действительно широкие и интеллектуально стимулирующие дискуссии о том, что такое информация, были почти совершенно неактуальны для развития современных информационных технологий, что часто приводило к непредвиденным впечатляющим прорывам, в то время как теоретики все еще обсуждали последствия более ранних, столь же неожиданных достижений.

Обсуждение концептуальных требований обработки информации в процессе исторического источнико-ориентированного исследования имеет смысл, следовательно, только если мы можем указать с некоторой точностью, где технология, которой мы располагаем сегодня, препятствует или ограничивает рассуждения в пределах соответствующих дисциплин. И хотя полное техническое решение такой адаптированной технологии является действительно сложной задачей, указать, где именно необходимы исследования и технические разработки, вполне выполнимо. Определены четыре такие области для исследований.

Искажение Вивера

Выше мы утверждали, что во время популяризации работы Шеннона [*Weaver,* 1949, p. 25] в информатику было введено ложное убеждение, что улучшение обработки данных автоматически улучшит обработку информации и даже знаний. В результате у нас есть формализмы для обработки данных, которые привели к доступности многих повсеместных решений для самых общих уровней программирования данных. Однако решения для технической обработки проблем на двух более высоких концептуальных уровнях обычно сопряжены с особенностями частичных решений, которые часто несовместимы.

Чтобы улучшить эту ситуацию, мы предлагаем исследование в двух направлениях. С одной стороны, реализация предложения Девлина [*Devlin,* 1991] по математике информации. С другой стороны, переосмысление предположения о том, что информационные системы обрабатывают статические структуры данных с помощью динамических алгоритмов в направлении решения, в котором структуры появляются только в виде моментальных снимков в состоянии постоянно работающих алгоритмов, основанных на обобщении мной концепции Лангефорса [*Langefors,* 1973; *Thaller,* 2009, 345ff.].

Двоичная ошибка

То, что цифровые компьютеры построены на двоичных числах, привело к неправильному пониманию того, что все программирование обязательно должно быть также двоичным. В действительности сегодня практически во всех областях применения информационных технологий существуют решения для применения более совершенных логических моделей, которые в целом называют «нечеткими» в признании оригинальной работы Заде [*Zadeh,* 1965, 1975, 1978, 1999]. Тем не менее такие приложения опять же склеены только вторично на бинарных стандартных технологиях и создают изумительное множество идиосинкразий.

Поэтому мы выступаем за интеграцию обобщенных решений в высшие языки программирования для трех категорий задач: (1) нечеткость в более узком смысле, т.е. невозможность придать четкое истинное значение утверждению; (2) присущая неточности семантическая концепция, как у «стариков»; (3) элемент, который концептуально является скалярным, но выходит за рамки наших текущих типов данных. Например, цена товара, для которой у нас нет точного значения, но есть минимум и максимум, плюс, возможно, намеки на распределение точек данных между ними.

Тупик Хомского

Обработка текстовых данных компьютерами находилась под сильным влиянием концентрации на синтаксисе, который сыграл важную роль в первые годы вычислений, но в настоящее время является препятствием для лучшей интеграции семантической обработки в обобщенные технические решения, хотя поддержка семантического содержания текстов – это главное требование для исторических исследований.

Альтернативный выход – представление Лакоффа и Джонсона о том, что понимание основано на метафорах [*Lakoff,* 1980] – становится незамедлительно интуитивным для историка, который пытался просмотреть значение, скрытое за историческими текстами. И идея о том, что возможность создавать ассоциации между различными понятиями, смешивать понятия [*Fauconnier,* 2003] на самом деле является особенностью, отличающей человеческий разум гораздо более фундаментально, чем «я-язык» (I-Language) и универсальная грамматика (Universal Grammar) [*Isac,* 2008], очень убедительна для того же историка.

Общая техническая поддержка для таких моделей требует поддержки классов графов в информатике, которые в настоящее время отсутствуют.

Ошибка разметки

Языки разметки не являются центральным предметом информатики. Однако многим гуманитариям они кажутся квинтэссенцией так называемых цифровых гуманитарных наук. На мой взгляд, у них на текущий момент при использовании разметки в обработке исторических документов имеется два методологических недостатка.

С одной стороны, встроенная разметка скрывает разделение представления источника и интерпретации его содержимого, поскольку обе задачи неразрывно смешаны в современных стандартах кодирования, особенно в TEI. С другой стороны, в настоящее время существует полное разделение встраивания интерпретаций в текст с помощью разметки и применения интерпретаций ко всем другим типам данных на основе внешних аннотаций.

Для решения этих двух проблем требуется техническая поддержка для последовательного решения тупиковой разметки, применимой ко *всем* типам данных. Кроме того, для поддержки текстов, которые были переданы нам в двух или более частично противоречивых версиях, мы должны заменить инструменты для обработки текста, основывающиеся на предположении, что все тексты являются линейными строками, более общим решением, реализующим их как графы.

Также кратко упоминаются несколько других улучшений современной технологии.