New Developments in Electronic Publishing
AMS/SMM Special Session, Houston, May 2004
ECM4 Satellite Conference, Stockholm, June 2004
p. 119-135.

Remarks on a Publication-Based Concept of Information

Wolfgang Lenski

Department of Computer Science, University of Kaiserslautern
P.O. Box 3049, D-67653 Kaiserslautern
e-mail: lenski@informatik.uni-kl.de

Abstract. The concept of 'information' is one of the most fundamental ones in computer science. Nevertheless it has not undergone a clear foundation that may compete with foundations in other scientific fields, especially in mathematics. Instead, there are many different approaches found in the literature which are specifically taylored to problem spaces. Since a general solution of the foundational issue seems to be out of focus at the moment, we concentrate in this paper on questions for a conceptualization that is centered around an understanding in the context of extractions of material out of publications as stored in publication databases maintained by publishing houses or database providers like Zentralblatt MATH.

1. Introduction

According to Thomas Kuhn's famous theory of scientific revolutions [34], it is a significant phenomenon of scientific developments that there are episodes in which 'normal science' just develops by increasing and cumulating achievements on topics of which a community shares an understanding and is especially interested in. It is then a common behavior of 'normal science' that problems occurring find their solutions within the conceptual and methodological framework of the paradigm which in turn gives raise to subsequent investigations. In this sense a scientific field just develops for some times. This was the situation in logic for more than 2,000 years starting with Aristotle. Even Kant in the 18th century was convinced that logic in the shape of the Aristotle's syllogisms has reached its final stage not being able of further development (see [32, B VIII]).

But when more severe problems show, deeper insight into the objects under consideration are gained, and new developments transcent the agreed stock of methods with which the

previous problems have been tackled, there evolves an increasing demand for clarifications and justification of the basics of the underlying paradigms. This demand aims at a specification of the fundamental concepts along with a clarification of their basic properties which is finally complemented by a methodological reflection. It is worth noting in this context that the requirement of a methodological reflection in this sense has even been posed for the philosophical system of Immanuel Kant as a whole; see [26].

In mathematics a considerable amount of investigations of that kind is subject of so-called *meta-mathematics* determining the allowed construction methods along with investigation of the (principal) limitations of a given methodology. It is worth mentioning – and it is a special peculiarity in mathematics – that meta-mathematical considerations can be developed inside the system itself, i.e. within logics. This is especially a consequence of the fact that via coding essential parts of meta-mathematical considerations may be dealt with within the framework of logic itself.

And then there are non-cumulative developmental episodes science in which an older paradigm is replaced in whole or in part by an incompatible new one or a new field of activities is envisaged. A change of paradigms is often the result, if the problems that have been faced appear unsurmountable, lack a plausible justification, or show unforeseen behavior which then results in a scientific crisis. It then causes an even increased need for foundational issues.

The foundational crisis in mathematics in the late 19th century reinforced by the detection of paradoxes is an example for such a scientific development. The subsequent revision installed a new paradigm, under which logic separated from philosophy and – initiated by Frege's *Begriffsschrift* [23] – finally became a formal theory of *truth* and *proof*. This change of paradigms turned out to be highly successful and finally led to a new foundation of mathematics based on logics. In the sequel logic even became 'the' model for justified foundations, not only in mathematics.

In the field of computer science we may identify a situation in which a clarification of the central concepts is increasingly demanded. It is broadly agreed upon that computer science has not yet reached such a satisfactory foundation according to these fundamentals. Some special subfields may have inherited foundational justification though by explicitly relying on external fields such as 'complexity theory' on logic or are rather tied to engineering, respectively. It is, however, especially true for one key subpart: information science. However, we end with rather structural dependencies for a justified model for retrieval – and no exact model for its behavior!

Even a short occupation with possible meanings of the term 'information' will immediately bring to light that very different and wide-spread understandings of the concept of information have evolved. "There is no accepted science of information" so the opening remark in the preface of the monography of Barwise and Seligman [3]. And as a broadly accepted *science* of information is not envisaged by now, a commonly acknowledged *theory* of information processing is out of sight as well. Compared to mathematics we are presently in a 'pre-Tarskian' stage of development. It is the intention of this paper to exhibit the grounds from which a theory of 'information' finally may be expected to evolve.

2. Information and computer science

Development of 'normal science' character happened in the sub-field of computer science where the question of storing and (efficiently) regaining previously stored material was eventually complemented by the question of exploiting the stored material to solve problems a person comes across. In this sense the original meaning of 'retrieve' as simply 'finding again'[1] was transcended and the field of *information retrieval* finally evolved. This development has essentially been influenced by the work of Gerald Salton in connection with the SMART-project [51] which marked the beginning of a new field in computer science.

Whereas the notion *information retrieval* already hint at some more advanced intentions, the underlying technique may rather be characterized as *document* retrieval where the system presents documents to the user. The documents in this sense just being hints to the material; it remains the task of the user to identify the specifically interesting content inside the documents and to extract the wanted information out of it. This constitutes an essential gap between the system's abilities on one hand and the user's information need on the other hand. It remains the main challenge in the field to design systems to serve the information need directly. From a systematic point of view, however, this would amount to an implementation based on a (formal) specification of information – an attempt that is out of sight at the moment. As a result there remains a systematic gap between the system's design principles and the user's intentions.

To compensate this deficiency – and to maintain still the claim to ultimately serve the user's information needs – evaluation techniques have been presented which related the system's functionalities to the original task. In this sense the notion of *pertinence* was introduced being the property that 'assigns an answer to an information need' ([33]; cf. also [30, p. 151]). It especially emphasizes the relation to an interpretation of the concept of 'information' now as part of the *evaluation* process.

More practical perspectives were enabled when the explicit relation to the *information need* was cut and replaced by a reference to its manifestation. The respective notion of *relevance* being the property that 'assigns an answer to an information request' ([30, p. 151]) remained a subjective notion though and no absolute consensus on the concept was achieved ([43, p. 305]; see also the survey on the literature in [10]). But this still explicitly excludes an analysis of the *process* of a human to end up with the relevance judgement which remained subject to studies rather in the cognitive science's part of information retrieval than in the evaluation context.

Subsequent developments then showed a further transition to the very idea of *information* retrieval where the parts that specifically could contribute to the need may be presented to the information seeker by the system directly. This marked the beginning of a new paradigm in information retrieval which increased the demand for a solid and generally accepted foundation of the concept of 'information' or at least of a clear understanding of the process of the process of 'getting informed'.

Developments like 'passage retrieval', 'question answering' or 'novelty' (just to mention a few to stand for others; see for example the topics in [15]) would profit tremendously from a

[1] According to Webster's Third New International Dictionary [57], 'retrieve' etymologically goes back to a modification of Middle French 'retrouver' in the sense of 'to find again'.

justified and acknowledged procedure to identify relevant parts in a document by the retrieval system itself as opposed to the judgement of a human as they already implicitly implement a certain understanding of *information*. So a necessary specification of *information* beyond rather philosophical definition remains a main challenge in this field.

It is worth mentioning that a fully specified implementation of a well-justified concept of information would not need an evaluation procedure any longer as it would be enabled to present exactly those items which after an evaluation of the system's relevance judgement would turn out to be relevant, i.e. contribute to resolve the problem state.

The present situation can be compared to the situation in logic before Tarski in [55] has provided a criterion for truth complemented by a formal definition. 'Truth' has not been invented by Tarski and there is a long tradition of understandings of this concept even in mathematics. To identify the necessary reductions of the conceptualization compared to an everyday understanding and to bring it into a shape suitable for formal (or mathematical) treatment is an absolutely outstanding achievement that can hardly be overestimated.

But a conceptualization of 'information' is just one part of the problem. 'information' in our sense describes a *process* of a transition from a state of uncertainty to another state of less uncertainty in one area of interest (cf. the 'anomalous states of knowledge' (ASKs) in [6]) which is essentially performed by the acquisition and processing of *external material*. This is the view of a publication-based concept of 'information'.[2] It especially covers – but is not restricted to – the situation of publication-based information systems like *Zentralblatt MATH* [58] or the *Bibliography of mathematical logic* [9].

In this view 'information' and the origination of its material content constitute two different fields that have to be brought together seamlessly for the sake of the superordinate aim. As a consequence we must not neglect the internal constitution of the objects providing material. Hence a publication-based conceptualization rules out all the approaches mentioned above; a theory of information in the sense we are interested in definitely cannot rely on these as all these formalizations concentrate on some restricted understanding of 'information'.

[2]There are already some formalizations of the concept of 'information' (see also Section 4.1 for some more approaches and [36] for more detailed description):

- Shannon's theory of information ([52], [53]) may be considered as a theory of transmission which reduces the concept of 'information' to the decrease of uncertainty without considering any contents.
- Hintikka's pioneering paper of semantic information [28] presents a theory of information in the shape of a logical interpretation of probability as opposed to a statistical (or frequency-based) theory of probability. A logical interpretation focuses on possibilities to distinguish alternatives by means of their formal expressions of the logical language which is mostly done by calculating its length.
- The separation of information from necessary interpretation allows to study its functional behaviour. This is the program of [19] and the main task of the theory of information *flow*; see, e.g., [21], [22], [31].
- Although certainly inspired by the basic considerations of Shannon's theory of information, Gregory Chaitin [17], Ray Solomonoff, and Andrej Kolmogorov (see, e.g., [37]) developed a different view of information. Instead of considering the statistical ensemble of messages from an information source, *algorithmic information theory* focuses on individual sequences of symbols. As such it is developed as a mathematical theory of words and problems related to coding, complexity issues and pattern recognition.

What makes the problem of a foundation of this setting complicated is essentially due to their contextuality. On the one hand, the need for information arises from a problem space in some context. On the other hand a document develops a more or less coherent view on some topic thus establishing another context. This raises the question of compatibility of the contexts.

The context-dependency appears probably in its purest form in mathematics which also has developed 'views' even in the abstract world of mathematical objects: Compare for example classical 'Hilbert-style' mathematics with constructive mathematics in the sense of Brouwer! Theorems remain dependent on their commitment and cannot be transplanted into the other context, respectively. But even beyond that the given results depend on definitions that may vary, and general commitments have to be observed like "For the rest of the book assume that ..." which makes it impossible to extract material just at the place where it occurs. Hence the challenge is two-fold: to present

- a foundation of the concept of 'information'
- a corresponding conceptualization of 'document'

that in addition meet the requirements related to contextuality.

Beyond an isolated treatment of these topics, a solution for a publication-based concept of 'information' as a whole must also address questions of consistency and coherence of the problem space and the document's contents. It is indispensable that the problem space and the suggested solution share the same understanding. This consistency – which only admits the exploitation of the material provided by the document – must then be guaranteed in a second step where special adaptation may be involved. But this is a downstream question that can only be addressed after a successful foundation of the two constituents.

Whereas the document's part has already been worked out to some extent (see 5), an broadly acknowledged foundation of 'information' it not in sight. We thus concentrate in this paper rather on questions concerning conceptualizations of 'information'. The first idea in this situation is certainly to use logic for the modeling purposes as it provides a formalized concept of a 'model'. So let us throw a closer look at the suitability of logic for this purpose.

3. Principles of logic as a modeling framework

After its paradigmatic turn in the 20th century, logic became 'the' model for modeling purposes in the sciences and even other disciplines as for example in analytical philosophy. However, to understand the principal reach of the representation potential of the mathematical (i.e. logical) methodology we must first sketch the philosophical background on which this methodology finally relies. This will throw a light on its suitability for the wanted (formal) specification of the concept of 'information'.

3.1. Philosophical background

The basic situation is comparable to the situation in logic concerning the concept of truth where Tarski [55] has explicitly referred to foundational philosophical positions, namely the semiotic approach as the philosophical ground for his pioneering theory of truth. This means that he explicitly wanted to prevent his approach from being some kind of ad hoc or based on

a specifically taylored background. In this sense a philosophical reflection seems inevitable and indispensable at the same time.

Our approach is committed to the semiotical theory in the tradition of the pragmatic philosophy of C. S. Peirce [50] as well. While even a rough outline of the philosophical insights would be out of the scope of this work, we will just sketch the very basic principles of this approach that may reveal the influences to the theory of information we are interested in. More careful introductions into the field may especially be found in [29], [49], [20], [18], [46], or [24]; for a more detailed discussion with special focus on the topic we are interested in cf. [36].

According to the philosophical tradition originating from the (idealistic) position of Kant, knowledge is composed via perception and subsumption of concepts under categories and – as Peirce especially emphasizes – is performed by individuals and results in an orientation in the 'world'. This is seen as a process in which first concepts are formed and finally knowledge is established. These concepts are closely tied to indended possible actions by that very person. This is expressed by the *pragmatic maxim* [50, 5.402].

Now what initiates our awareness about phenomena are *signs* which are considered as the very basic constituents of all epistemic processes. Hence this view essentially demands a theory of signs along with a theory of understanding and interpreting these which is developed in the field of semiotics.

Understanding may now be characterized as a principally uncompleteable process of the effect of signs for an interpretant. This process is called *semiosis* (cf. [50, Paragraph 5.484]). It basically implies a triadic understanding of signs as claimed by the categories [50, Vol. 8, Paragraph 328]: A sign is a First which stands in a genuine triadic relation to a Second, its object, and a Third, called its interpretant (cf. [14, 99]).

According to the pragmatic semiotics we have to deal with signs, concepts, and objects together with their relations. Since the main focus of this paper is not on problems of epistemology, but to base our theory on clear philosophical insights and to investigate the necessary steps for a methodologically well-justified theory of information we have to bridge the gap between the philosophical considerations and conceptualizations that can actually constitute a foundation for information system. This requires a transformation of principles of rather epistemic nature into conceptualizations that admit a more practical treatment. Such a transformation may be grounded on a purely analytical reinterpretation of this setting by Morris (cf. [45]). Accordingly, semiotics as the general theory of signs has three subdivisions [44] (see also [16]):

(1) *syntax*, the study of the formal relations of signs to one another,
(2) *semantics*, the study of the relations of signs to the objects to which the signs are applicable,
(3) *pragmatics*, the study of the relation of signs to interpreters.

3.2. A semiotic view of information

In this section we will show that the semiotical approach provides a theoretical basis for an unified view on data, knowledge, and information. The common basis of these concepts is provided by the abstract concept of a *sign* which is an ontological unity. But signs in

an epistemic context may also be subject to analytical considerations according to Morris' semiotical dimensions. This results in the following conceptual coordination (see also [36]):

Data. *Data* denotes the *syntactical* dimension of a sign.

In this sense *data* denotes the organized arrangement of signs with emphasize given on the structural or grammatical aspect only. Moreover, just arbitrary arrangements of signs are not considered as data, because data are always derived from an understanding of a part of the world. Accordingly, data are not interpreted, but interpretable. This requires the usage of a commonly understood organizational form, a grammar, which is then endowed with a shared interpretation within a community.

Knowledge. *Knowledge* denotes the *semantical* dimension of a sign.

The analytical abstraction contained in semantics explicitly abstracts from persons that actually ascribe its contents. What counts as knowledge is solely the view that results from neglecting the amount of individual contribution to the abstracting process. However, this very process may well be subject to inquiries itself. These investigations are necessarily part of the procedure to *account for* semantical relationships. It gives raise to several dimensions or degrees of justification that may be distinguished.

The ideal would certainly be that such an abstraction process indeed can be performed in an absolute sense. This would be the ideal of a pure semantic characterization and is in general the pretension of *truth*. Mathematical results are mostly considered of being of that kind, i.e. independent from the person that demonstrated its validity: There should be no *subjective* mathematics that might differ from person to person. Such an approach has for example been tried to establish by Kant when relating mathematics to 'a priori forms of perception'. This is at the same time a justification for a foundation of mathematics on the basis of truth as provided by logic.

But this essentially requires that every person shares the same conviction on which these abstraction processes can ultimately be based. As this is certainly not the case, we end up with some sort of intersubjective commitment ("knowledge") which thus is authoritative to that part of the community only which shares the underlying presuppositions. Even Tarski's semantic theory of truth [55] in mathematics has not experienced general acceptance as a universal methodology and remains bound to Hilbert-Tarski-style of mathematics; see, e.g., constructive mathematics in the sense of Brouwer. The point is essentially a methodological question of admitting some general principles or not (see, e.g., [56]).

Finally, such an abstraction process may well be only subjectively acknowledged resulting in only personal knowledge which does not comprise a universal claim of justification.

Information. *Information* denotes the *pragmatical* dimension of a sign.

Being of pragmatical dimension, information demands an interpretant to perform an interpretation. Information is thus bound to a (cognitive) system to process the possible contributions provided by the sign (the data) for a possible action. Moreover, information inherits the same interpretation relation as knowledge with the difference that the latter abstract from any reference to the actual performance of the interpretation whereas the former just emphasizes these. Altogether, this epistemic position clearly shows that knowledge and information are closely tied together.

3.3. Representation in logic

The scientific ideal in its major form is certainly mathematical logic. Its predecessor, Aristotelian logic, has remained rather a philosophical discipline. It presented a (formal) theory of correct thinking but remained bound to philosophical disputation as the sole ground for justification.[3] As such it was not suitable for foundational issues in the sciences mostly because of its lack for concept specifications which had to be provided from outside the very theory.

It was in the midth of the nineteenth century when a scientific break-through marked a change of the main paradigms. This was essentially due to the work of Frege whose *Begriffsschrift* [23] marked the beginning of a new area in logic. A new foundation of logic based on *proof* and *truth* instead of philosophical consideration was finally completed when Tarski [55] presented his truth-criterion which was supplemented by a formal theory of truth.[4] Moreover, as a consequence of the famous results of Frege, Gödel, Tarski, Kleene – just to mention a few to stand for many others who also contributed to this question – logic is at the same time 'meta-logic' able to reflect the basic foundational problems, e.g. consistency, inside its own framework. As a result mathematical logic

- provides an abstract representation methodology for conceptualizations
- admits the representation of conceptual dependencies
- incorporates a completely formalized justification procedure
- reflects methodological issues of its own field.

This turns logic into a foundational science and makes it especially suitable for the representation requirements in mathematics which then inherits its foundational status from logic. But there are other applications of this foundational framework as well as for example in the philosophy of science (cf., e.g., [54] or [47]) or in some subfields of computer science like complexity theory. Hence logic is 'foundational science'. So naturally the question raises about its suitability in information science.

We will argue, however, that logic does only show limited applicability in this field – and 'information science' rather is a scientific discipline of different scope which cannot be reduced to logic in the sense of mathematics. This is illustrated by some consequences of the application of logic for modeling purposes.

A first indication is provided by Hintikka's so-called "logical omniscience" problem [27]: An agent is *omniscent* if he knows all true consequences of a given statement. But according to the foundation provided by Frege and Tarski logic is just constituted as a theory of truth and proof. Hence every theory based on (logical) derivations is necessarily omniscent: a theory includes (at least theoretically by making use of the deduction facilities provided by logic in general) all its consequences! This is just a heritage from the very conception of logic. On the other hand no person is omniscent! Otherwise outstanding problems could immediately be solved: Present for example the axioms of Zermelo-Fraenkel set theory to a person, and he/she – being omniscent – would immediately be able to tell whether these

[3] See for example the textbook [2] which had constituted a most influential work in the seventeenth century.

[4] It must be mentioned at this point that in addition to Frege and Tarski the outstanding work of Gödel (see esp. [25]) also contributed essentially to this foundational program insofar he provided a fundamental clarification of the possible reach of the concept of 'proof' within this framework.

axioms are consistent, i.e. do not admit the derivation of a contradiction, since that person would oversee all true statements being provable consequences of these axioms.

This last example demonstrates that 'information' is *not* truth-preserving, but every (reasonable) theory based on proof and truth is necessarily truth-preserving. This shows that the required theory of 'information' cannot be developed inside the framework of logic without additional adjustments that try to compensate its theoretical deficiencies to some extent.[5] From a systematic point of view these must be understood as attempts to amend some subordinated issues instead of acknowledging the principal inadequacy of the methodology itself.

The previous philosophical investigations, however, which as such transcend the reach of a field's own methodologies show us the deeper reason for this inadequacy: it is actually a foundational issue! So let us then analyse the limitations of logic from a fundamental point of view.

Tarski has explicitly chosen a *semantic* foundation of the concept of truth. According to our semiotical analysis, this abstracts from the interpretant in the semiotical analysis and hence relates truth to knowledge. This seems to be a canonical decision, and it is doubtful whether Tarski actually would have had another choice. Especially, this decision meets all the requirements that are commonly associated with mathematical (and logical) knowledge: The results of mathematics have always been thought of not only as being true but also as being *necessarily* true. In the sense of Aristotle this means *true at all times*. It especially implies that the results are independent of the person who establishes them. Moreover, their validity only relies on the presuppositions and the results are thus insensitive to their context (though dependent on the basic principles separating, e.g., Hilbert-style mathematics from constructive mathematics). All this is in perfect conformity with abstract requirements for a conceptualization of 'truth' that one may pose and turns 'truth' into a concept that is (at least in principle) suitable for foundational purposes. So logic is a theory explicitly taylored for knowledge. As such it is not problem-induced; this part is left to personal interests while logic is only concerned with the *results* of corresponding investigations.

On the other hand, this choice is in explicit contrast to the concept of *information* which is inherently of pragmatic nature, i.e. it is not only dependent of a person but also of the situation he/she is currently in. Hence it is clear that without special adaption which restrict its abilities in some sense logic is *not* able to provide the methodological framework for a conceptualization along with a subsequent theory of 'information'.

Hence a justified foundation must be based on something else. To start with, we will first exhibit the epistemic status of the concept of *information*. This should then determine the abstract properties of 'information' constituting the indisposable basis for a subsequent modeling.

4. Towards a conceptualization of information

The philosophical foundation has provided the epistemological background which leaves open the question of subsequent specifications. It is the task of the sciences to establish specifica-

[5]There have been attempts to restrict deduction procedures by exerting control over the deduction process such as to limit the number of steps.

tions that are consistent with the philosophical considerations and concretize aspects for their special purpose. In the following section we will investigate the contexts of such realizations.

4.1. The fundamental equation of information science

There have been attempts to formalize conceptualizations of 'information' in information science. In particular, Mizzaro's theory of information in [41], [42] provides an interesting approach to define information which essentially tries to capture the result of an information seeking process via its traces in the very result of the process, i.e., via the effect it has on given structures. It may thus be characterized as a post-process definition of information as opposed to in-process definitions which seek to capture the very process of getting informed. As we are rather interested in an insight into the processes that result in transformations of given structures which may be characterized as 'being informed', we do not follow this line.

In general, it seems inevitable that certain reductions of the conceptual complexity of a socio-cultural understanding of the primary concepts have to go with such attempts. The semantic conception of 'truth' does definitely also not capture all facets of the understandings of the concept in real life! Instead, it is specifically tailored to capture the fundamental properties that are needed to build mathematics on it. All approaches towards a formalization of 'information', however, are far from constituting a formal system in the spirit of the theories described in this section – and a respective theory is out of sight at the moment.

The most promising starting point in view of admitting possible specifications of the basic symbols in terms of the cognitive sciences is certainly the famous 'fundamental equation' of information science which has been discussed in [11] and successively refined until reaching its famous form in [13]:

$$K[S] + \Delta I = K[S + \Delta S] \tag{1}$$

This equation

> *... states in its very general way that the knowledge structure $K[S]$ is changed to the new modified structure $K[S+\Delta S]$ by the information ΔI, the ΔS indication the effect of the modification.* ([13, p. 131])

Brookes has already attached some background ideas and annotations to statement (1) (cf. [13, p. 131]). Especially, he points out that this equation must not be misunderstood as a simplification but as a *representation* of more complex insights. In this sense it is rather meant to transscript some general insights into the 'nature' of the mutual dependency of the concepts of knowledge and information.

Brookes' 'fundamental equation' provides the most abstract formal specification of the *interaction* of data, information, and knowledge and has experienced a broad acceptance in the information science and retrieval community – but it does not represent a *definition*. Our considerations are now meant to provide a conceptual clarification of the significance of the fundamental equation beyond the original background. The respective clarifications are at the same time the first step for any further modeling insofar they determine the abstract properties of and interrelations between the concepts that must be considered. This will result in a specification of the *principles* behind (1) in a theoretical framework (for a discussion see also [30]) that should finally give raise to a corresponding definition.

4.2. Principles underlying information

In this section we summarize properties along with the functionalities found in the literature that contribute to a unified view and have to be modeled by a theory of information. The principles (1) to (0) will constitute a system of prerequisites for any subsequent theory of information that claims to capture the essentials of a publication related concept of information.

(1) Information is a difference that makes a difference. [5]
(2) Information is the values of characteristics in the processes' output. [38, p. 256]
(3) Information is that which is capable of transforming structure. [7]
(4) Information is that which modifies [...] a knowledge structure. [12, p. 197]
(5) Knowledge is a linked structure of concepts. [13, p. 131]
(6) Information is a small part of such a structure. [13, p. 131]
(7) Information can be viewed as a collection of symbols. [40]

A few remarks to these are meant to complement the pure citation (for a more discussion of these topics see [36]). Their relation to the 'fundamental equation' will be sketched in short:

(1) denotes the overall characterization along with the functional behavior of information. It refers to the Δ-operator in the 'fundamental equation'.
(2) expresses that there is a process involved whose *results* ΔI are the constituents of information. This is the part *information retrieval* especially focuses on.
(3) specifies the abstract 'difference' as a transformation process resulting in $K[S + \Delta S]$. At the same time it hints at some regularities ('structure') that must be present to impose influence upon.
(4) shows on what information causes effects, namely on knowledge (structures): $K[S]$
(5) characterizes the necessary internal constitution of such background *structures* (i.e. nothing being amorph) and especially provides a further condition on 'knowledge' (structures) to be able to admit effects at all. It also implies that the extraction of subparts is actually feasible and constitutes sense.
(6) relates the internal constitution of information units to the internal constitution of knowledge structures. It states that certain compatibility conditions must be fulfilled in order to be able to impose effects.
(7) specifies the internal constitution of the *carriers* of information. Moreover, it gives the contents which in principle would admit a formal treatment. According to the semiotic approach these must be *signs*.

The principles we found complement the formal statement and the abstract discussion as well. They provide a pre-formal intermediate layer between pure philosophical considerations and thus prepare for a *definition* instead of a description of their interaction.

Beyond a determination of the interaction of 'data', 'knowledge', and 'information', a definition of 'information', however, must in addition reflect its semiotical status. This is not specified in the principles given above. Being of essentially pragmatic nature, information requires a concrete situation in which a previously unknown part of a knowledge structure

has to be acquired. This involves an aspect of *causation*. This makes information 'problem-driven': it is a problem that has to be resolved what initiates a differentiation process as specified in principle (2).

(8) The emergence of information is problem-driven.

Observing all this we present a working definition for a publication-related concept of 'information' which opens a perspective for future formalizations:

Definition 1 *Information is the result of a problem-driven differentiation process in a structured knowledge base.*

It is worth mentioning at this point that this definition leaves open the question of a problem-formulation that can be processed inside the knowledge base (the document model in our setting). An envisaged formalization must especially include such a language.

Finally, what can be searched for depends on the offerings of the 'knowledge base'. In the context of publications it is the document model with its internal structure which is considered in the following section.

5. The document

The concept of a 'document' as considered in this paper is in its most abstract form an object (e.g. an article or a book) which is described as an entity which

- is devoted to a topic
- develops a coherent view on that topic
- is created and composed out of individual units (which may be called *semantic units*) under a unifying idea and by utilizing of work at hand.

The notion 'coherent view' implies that there are constraints which are meant to guarantee a certain consistency of the material contained therein as opposed to arbitrary collections of units of a given form. Moreover, in order to be used in a setting of information processing, at least the subparts ("semantic units") of which the document is composed must be isolated and automatically extractable from the document. This implies that it must have an underlying organizational structure admitting (direct) access to its units. Moreover, this definition does allow an inductive generation process of documents based on each other (for some restrictions see below). 'Publication-based' now can been seen as a requirement which constitutes and is able to handle context for the units taken out of the corresponding documents.

Please observe that this characterization excludes items of some 'dadaistic' nature, i.e. which is produced in an arbitrary sense. This characterization may well exclude some kinds of productive work but as we are mostly interested in scientific works we may impose this restriction which in effect constitutes a more challenging property.

For such a structure there are already promising approaches for the desired functionality in sight. The *Document Object Model* (DOM)[6] is a platform- and language-neutral interface that will allow programs to dynamically access and update the content, structure and style of documents. As a consequence, the document can be subject to further processing

[6]See `http://www.w3.org/DOM`

which especially includes isolation of subparts for further processing. In combination with an underlying XML-structure it may admit at least a part of the desired functionality.

This leaves open the problem of contextual dependencies. In mathematics, this means that there has to be observed (besides other presuppositions) at least whether a proposed solution is bounded to classical mathematics, essentially relies on constructive conceptualizations, or is universally valid in case its justification prodecure ('proof') does not involve core principles.

The EU-funded project TRIAL-SOLUTION addresses the description of a document as given above in a framework that supports the free generation of documents by observing such contextual dependencies of the parts to be composed into the comprehensive document. This is performed via an appropriate set of meta-data descriptions to control the generation that can be processed by a corresponding functionality. It also admits the automatic enlargement of the composed document by adding prerequisites that have not been included so far but whose presence is desirable in the given context (see for example the composition of a textbook in a field in mathematics one is not fully acquainted with). Note that such a document may only consist virtually by means of a construction procedure that determines how to build the document under consideration out of its parts.[7]

This process necessarily requires a control structure which (in analogy to set theory) may be characterized as *well-founded*, i.e. does not lead to a circle via some sort of self-reference. It goes beyond the DOM approach insofar as it observes intellectual copyrights that may be associated with the individual units as well as with the finally composed document. To serve this purpose, it has introduced a meta-data description procedure to account for these rights which in addition fully conforms with usual citation practise (see [35]).

For the functionality envisaged in this paper it would then just remain to compare presuppositions of each (generated) document to guarantee consistency. This remains a major challenge, though, on the document processing part of information retrieval.

6. Outlook

We have shown that the 'fundamental equation' of Brookes indeed incorporates – in coded symbolic form – essential determinants of the concept of 'information' in a publication-related context. Rather philosophical investigations have exhibited the concepts that are necessarily connected with it. At the same time this admits a concentration on the very concepts that may contribute to a solution and rules out some extensions like 'wisdom' as suggested by [1] for which there is no place left in the systematization provided by the semiotical approach in the succesion of Peirce.

In this sense this paper contributes to the general research program initiated by Brookes [13, p. 117] when stating that

> the interpretation of the fundamental equation is the basic research task of information science.

We then have discussed perspectives for a corresponding conceptualization of 'information' on the basis of a conceptual 'ontology' which may be considered as establishing (sort of) a

[7] For more information see the project's home page `http://www.trial-solution.de`

conceptual pre-axiomatization. The philosophical background prevents it from just having produced an ad-hoc suggestion. Hence it constitutes a scientific perspective for future work in this field to develop a foundation which is in full accordance with the abstract requirements we have established.

What is still missing, though, is a precise (hopefully *formal*) foundation of the key notion 'information' which complements a definition with a formal specification as it has been provided for the concept of 'truth' by Tarski. However, the necessary steps towards such an aim have been sketched. It is to be hoped that such a foundation will finally evolve providing the field if information science with a well-justified foundation.

References

[1] Ackoff, R. L.: *Transformational consulting*. Management Consulting Times **28**(6) (1997).

[2] Arnauld, Antoine; Nicole, Pierre: *La Logique ou L'Art de penser*. 6th extended ed. 1685. German translation as "Die Logik oder Die Kunst des Denkens" by Christos Axelos. Wissenschaftliche Buchgesellschaft, Darmstadt, 2nd ed. 1994. XXVIII, 364p.

[3] Barwise, J.; Seligman, J.: *Information Flow: The Logic of Distributed Systems*. Cambridge University Press, Cambridge 1997. Zbl 0927.03004

[4] Bateson, G.: *Steps to an ecology of mind*. Paladin, St. Albans, Australia, 1973.

[5] Bateson, G.: *Mind and nature: a necessary unit*. Bantam Books, 1980.

[6] Belkin, N. J.; Oddy, R. N.; Brooks, H. M.: *ASK for information retrieval: Part I. Background and theory*. Journal of Documentation **38**(2) (June 1982), 61–71.

[7] Belkin, Nicholas J.; Robertson, Stephen E.: *Information Science and the Phenomenon of Information*. Journal of the American Society for Information Science 1976, 197–204.

[8] Belkin, Nicholas J.: *The Cognitive Viewpoint in Information Science*. Journal of Information Science **16** (1990), 11–15.

[9] *Bibliography of mathematical logic*. Electronically accessable via `http://www-logic.uni-kl.de`. Printed version Müller, G. H.; Lenski, W. et al.: Ω-*Bibliography of Mathematical Logic, Vols. I–VI*. Springer-Verlag, Heidelberg 1987.
Zbl 0632.03002 Zbl 0632.03003 Zbl 0632.03004 Zbl 0632.03005

[10] Borlund, Pia: *The Concept of Relevance in IR*. Journal of the American Society for Information Science and Technology **54**(10) (2003), 913–925.

[11] Brookes, Bertram C.: *The Fundamental Equation in Information Science. Problems of Information Science*. FID 530, VINITI, Moscow 1975, 115–130.

[12] Brookes, Bertram C.: *The Developing cognitive viewpoint in information science*. In: de Mey, M. (ed.): International Workshop on the Cognitive Viewpoint. University of Ghent, Ghent 1977, 195–203.

[13] Brookes, Bertram C.: *The Foundations of Information Science. Part I. Philosophical Aspects*. Journal of Information Science **2** (1980), 125–133.

[14] Buckler, J. (ed.): *Philosophical writings of Peirce*. Dover, New York 1955.

[15] Callan, Jamie; Cormack, Gordon; Clarke, Charles; Hawking, David; Smeaton, Alan: SIGIR 2003. The Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, 2003.

[16] Carnap, Rudolf: *Foundations of Logic and Mathematics.* University of Chicago Press, Chicago 1939. Zbl 0023.09701

[17] Chaitin, Gregory J.: *Algorithmic Information Theory.* Cambridge University Press, Cambridge 1990 (3rd ed.). cf. 2nd ed. 1987 Zbl 0655.68003

[18] Deledalle, Gérard: *Charles S. Peirce: An Intellectual Biography.* John Benjamins, Amsterdam 1990.

[19] Dretske, Fred: *Knowledge and the Flow of Information.* MIT Press, 1981.

[20] Fisch, Max H.: *Introduction to Writings of Charles S. Peirce.* Indiana University Press, Bloomington 1982.

[21] Floridi, Luciano: *Outline of a Theory of Strongly Semantic Information.* Submitted; also available at `http://www.wolfson.ox.ac.uk/~floridi/pdf/otssi.pdf`.

[22] Floridi, Luciano: *Is Information Meaningful Data?* Submitted; also available at `http://www.wolfson.ox.ac.uk/~floridi/pdf/iimd.pdf`.

[23] Frege, F. L. Gottlob: *Begriffsschrift, eine der arithmetischen nachgebildete Formelsprache des reinen Denkens.* Nebert, Halle 1879. JFM 27.0045.02

[24] Gallie, W. B.: *Peirce and Pragmatism.* Penguin, Harmondsworth 1952.

[25] Goedel, Kurt: *Über formal unentscheidbare Sätze der 'Principia Mathematica' und verwandter Systeme I.* Monatshefte Math. Phys. **38** (1931), 173–198. Zbl 0002.00101

[26] Henrich, Dieter: *Systemform und Abschlussgedanke. Methode und Metaphysik als Problem in Kants Denken.* In: Gerhardt, Volker; Horstmann, Rolf-Peter; Schumacher, Ralph (Eds.): Kant und die Berliner Aufklärung. Akten des IX. Internationalen Kant-Kongresses. Bd I: Hauptvorträge. de Gruyter, Berlin 2001, 94–115. Also in: Information Philosophie, December 2000, 7–21.

[27] Hintikka, K.; Jaakko J.: *Knowledge and Belief.* Cornell University Press, Ithaca, NY, 1962.

[28] Hintikka, K.; Jaako J.: *On semantic information.* In: Hintikka, K. Jaako J., Suppes, Patrick (eds.): Information and Inference. Reidel, Dordrecht 1970, 3–27.

[29] Hookway, Christopher: *Peirce.* Routledge & Kegan Paul, London 1985.

[30] Ingwersen, Peter: *Information Retrieval Interaction.* Taylor Graham, London 1992.

[31] Israel, David; Perry, John: *What is Information?* In: Hanson, Philip (ed.): Information, Language, and Cognition. University of British Columbia Press, Vancouver 1990, 1–19.

[32] Kant, Immanuel: *Kritik der reinen Vernunft.* Riga 1789.

[33] Kemp, D. A.: *Relevance, Pertinence and Information System Development.* Information Storage & Retrieval **10** (1974), 37–47.

[34] Kuhn, Thomas: *The Structure of Scientific Revolutions.* University of Chicago Press, 1962.

[35] Lenski, Wolfgang; Wette-Roch, Elisabeth: *Metadata for Advanced Structures of Learning Objects in Mathematics. An Approach for TRIAL-SOLUTION.* Available at `http://www-logic.uni-kl.de/trial/metadata_v2-0.pdf`.

[36] Lenski, Wolfgang: *Towards a Theory of Information.* In: Lenski, W. (ed.): Logic versus Approximation. Lecture Notes in Computer Science **3075**. Springer-Verlag, Heidelberg 2004. Zbl 1052.68005

[37] Li, Ming; Vitanyi, Paul: *An Introduction to Kolmogorov Complexity and Its Applications.* Springer-Verlag, Heidelberg 1997 (2nd ed.). Zbl 0866.68051

[38] Losee, Richard M.: *A Discipline Independent Definition of Information.* Journal of the American Society for Information Science **48**(3) (1997), 254–269.

[39] Machlup, F.: *Semantic Quirks in Studies of Information.* In: Machlup, F.; Mansfield, U. (eds.): The Study of Information, John Wiley & Sons, New York 1983, 641–671.

[40] Mason, Richard: *Measuring Information Output: A communication Systems Approach.* Information and Management **1** (1978), 219–234.

[41] Mizzaro, Stefano: *On the Foundations of Information Retrieval.* In: Atti del Congresso Nazionale AICA'96 (Proceedings of AICA'96), Roma, IT, 1996, 363–386.

[42] Mizzaro, Stefano: *Towards a theory of epistemic information. Information Modelling and Knowledge Bases.* IOS Press **12**, Amsterdam 2001, 1–20. Zbl 1004.68167

[43] Mizzaro, S.: *How many relevances in information retrieval?* Interacting with computers **10** (1998), 303–320.

[44] Morris, Charles W.: *Foundation of the theory of signs.* In: International Encyclopedia of Unified Science, Vol. 1, No. 2, University of Chicago Press, Chicago 1938.

[45] Morris, Charles W.: *Signs, Language, and Behaviour.* New York 1946.

[46] Murphey, Murray G.: *The Development of Peirce's Philosophy.* Harvard University Press, Cambridge 1961.

[47] Nagel, E.: *The structure of Science.* Hackett, Indianapolis 1979.

[48] Nyquist, H.: *Certain factors affecting telegraph speed.* Bell System Technical Journal **3** (1924), 324–346.

[49] Parker, Kelly A.: *The Continuity of Peirce's Thought.* Vanderbilt University Press, Nashville, TN, 1998.

[50] Peirce, Charles S.: *Collected Papers of Charles Sanders Peirce*, Vols. 1–8.
Hartshorne, Charles; Weiss, Paul (Vols. 1–6), Harvard University Press, Cambridge, Mass., 1931–1958. Vol. I: JFM 57.0038.02
Vol. II (in edition 1960 Vol. 1–6): Zbl 0173.00105
Vol. III and IV: JFM 59.0849.02
Vol. V and VI: JFM 62.1043.05
Burks, Arthur W. (Vols. 7–8) (eds.) Zbl 0081.00410

[51] Salton, Gerald: *The SMART Retrieval System.* Prentice-Hall, Englewood Cliffs, NJ, 1971.

[52] Shannon, Claude E.: *A Mathematical Theory of Communication.* The Bell System Technical Journal **27** (1948), 379–423, 623–656.

[53] Shannon, Claude E.; Weaver, W.: *The mathematical theory of communication.* University of Illinois Press, Urbana 1949. Zbl 0041.25804

[54] Suppes, Patrick (ed.): *The Structure of Scientific Theories.* Univ. of Illinois Press, Urbana 1977.

[55] Tarski, Alfred: *Der Wahrheitsbegriff in den Sprachen der deduktiven Disziplinen.* Anzeiger der Österreichischen Akademie der Wissenschaften, Mathematisch-Naturwissenschaftliche Klasse, **69** (1932), 23–25. JFM 58.0997.03

[56] Troelstra, Anne; van Dalen, Dirk: *Constructivism in Mathematics. An introduction.* Volume I, II. North-Holland Publ. Co., Amsterdam 1988. Vol. I: Zbl 0653.03040 Vol. II: Zbl 0661.03047

[57] Webster's Third New International Dictionary. Merriam-Webster, Springfield, MA, 2003.

[58] Zentralblatt MATH. Electronically accessable via `http://www.emis.de/ZMATH/`.

Received October 1, 2004