Kim H. Veltman

Challenges for a Semantic Web

"*Semantic Web Workshop 2002. Proceedings of the International Workshop on the Semantic Web 2002 (at the Eleventh International World Wide Web Conference), Hawaii, May 7, 2002,* Honolulu, Hawaii, pp. 16-22.
Position paper also published electronically at:
http://semanticweb2002.aifb.uni-karlsruhe.de/proceedings/Position/veltmann.pdf.
Reprinted in Cultivate Interative, Issue 7, June 2002
(http://www.cultivate-int.org/issue7/semanticweb/).

These ideas have since been developed in the author's "Towards a Semantic Web for Culture" (2004).

Abstract
The semantic web should be about the meaning of humans with all the richness of cultural and historical dimensions. This paper reviews three approaches to the semantic web, namely of the W3, Dublin Core and a small group within the AI community. It then suggests that a new kind of cultural semantics is needed in order to reflect the richness of human experience.

Categories and Subject Descriptors
Historical Semantics, Cultural Meaning

General Terms
Standardization, Theory.

Keywords
Semantics, Culture, History

---

1. Introduction

The semantic web[1] is analogous to motherhood and apple pie. Everyone agrees that it is a good idea. Semantic, as the Oxford English Dictionary tells us, has to do with meaning and everyone wants meaning.[2] As is so often the case when everyone thinks that they agree, it may be that the meaning of meaning is not as clear as it seems; that persons are actually speaking about different things, and that there is a danger that they are speaking past each other. This paper suggests that there are at least four approaches to the semantic web, namely that of:

1) World Wide Web (W3)
2) Dublin Core
3) a small group within the AI community
4) cultural semantics.

A brief survey of the four approaches is given. It is claimed that the first two approaches are correct but too narrow; that the third is misleading, while the fourth represents a direction full of challenges to which we should aspire.

2. Semantic Web of W3

At WWW7 (Brisbane, 1997), Tim Berners-Lee outlined his vision of a global reasoning web. At WWW8 (Toronto, 1998), he articulated the vision of a semantic web, whereby one can separate rhyme from reason: i.e. the subjective dimensions of art and poetry from the objective dimensions of logic, which is one definition of science. At one level, this is a direct continuation of the vision, which inspired Shannon, which itself grew out of the subject-object distinction that Cassirer[3] traced back to the Renaissance. In some senses it also goes back to the Greek debates about universals and particulars. In terms of the classical *trivium* of grammar (the structure of language), dialectic (the logic of language) and rhetoric (the effects of language), the emphasis of Tim Berners Lee on the logic of language reflects the concerns of dialectic in Antiquity.

In the vision of Tim Berners-Lee,[4] there is a great emphasis also on distinguishing the basic structure of content from the various forms in which it is expressed. In the *trivium,* this is the distinction between grammar (the structure of language) and rhetoric (the effects of language). There is corresponding attention to the *quadrivium.* Optimists will note that the makers of the World Wide Web (W3) Consortium are addressing all the questions of the ancient *trivium* and *quadrivium* such that all the potentials of the traditional seven liberal arts will soon be available in electronic form (figure 1). At the same time there is a danger in being overoptimistic and in being too easily satisfied. Separating rhyme from reason is useful. Creating a web which focusses only on reason at the expense of poetry may not be sufficient.

Logic is, of course, an excellent starting point. Tim Berners-Lee has a conviction, which can be traced back to early history of Oxford from which he comes, that logic is a way to separating the wheat of truth from the chaff of idle claims.

Grammar	Structure, Syntax[5]	Extensible Markup Language[6]	XML
Dialectic	Logic, Semantics	Resource Description Framework	RDF
Rhetoric	Effects, Style, Pragmatics[7]	Extensible Style Language	XSL
Geometry	Continuous Quantity	Mathematical Markup Language	MML
Arithmetic	Discrete Quantity	Mathematical Markup Language	MML
Astronomy	Applied Continuous Quantity	Astronomical Markup Language	AML
Music	Applied Discrete Quantity	Standardized Music Description Language	SDML

Figure 1. The seven liberal arts (*trivium*, *quadrivium*) and modern electronic equivalents.

Logic is universally applicable: it reflects the scientific spirit. It represents the dimension concerning which there ought, in theory, to be no debate. Logic has the added value that it can be very useful in the realm of transactions. If we can sort out which accounts are true and which false, this can help greatly the rise of e-commerce.

All this is excellent. Meaning, however, is about much more than transactions. Whereas the meaning in logic and science focuses on the universally true, meaning in the realms of culture typically focusses on what is nationally, regionally or locally unique. Science is in large part uni-lingual and uni-cultural. Culture is multi-lingual and multi-cultural. The solutions of science have become the models for our treatment of all domains of existence. Today when we search for a word on the Internet there is an implicit assumption that we are searching for a single meaning. For the realms of culture we need a semantic web, which allows us to discover differences in meaning in different places and at different times. We shall return to this in section 4.

3. Dublin Core

The W3 Consortium works closely with the Dublin Core (Metadata Initiative), which was inspired in part by the vision of Yuri Rubinsky (1994) for a metadata semantics.[8] This set out to identify a minimal set of universally applicable fields on which one could hope to gain international acceptance. These fifteen fields, known as the Dublin Core, were initially intended to describe web sites developed by persons without formal training in the principles of library cataloguing (e.g. MARC). In the eyes of some the Dublin Core has much grander applications in memory institutions. In any case it can serve as a very useful bridging device to connect otherwise heterogeneous resources. The Dublin Core initiative helps to reach agreement on matching effectively equivalent fields in different systems: a process which is alternatively called mapping, bridging, linking, creating crosswalks, walkthroughs or more generally interoperability. Interoperability of content is at least a twofold problem. There is interoperability of:

1) fields: i.e. we must agree that the field Author and Name are equivalent
2) meaning of the terms in those fields.

The initiators of the Dublin Core use semantics to refer to the definition or meaning of the fields (or elements). They deal with part one of the problem and this is very important. Without basic agreement concerning the fields there can be no sharing of

information and knowledge. In other words, qua fields/elements/containers we must first decide that Subject and Topic are equivalent. But interoperability of content entails a second part: qua meaning of terms in the fields we then need to agree that the subject/topic of car and the subject/topic automobile are equivalent.

In the case of car and automobile almost everyone will agree that the terms are equivalent. In the case of a word such as pasta, in Italy alone there are well over 60 definitions. In science, one internationally accepted definition of a term or word is all that is needed. By contrast in the realm of culture there is typically a definition at the international level and variants at the national, regional and local levels. Both the W3 and Dublin Core use science as a model. This approach based on logic and universals is excellent in the case of scientific knowledge, but is too narrow to deal with the particulars of multi-lingual, multi-cultural and historical cultural knowledge. For this we need a cultural semantics.

The authors of the Dublin Core and the W3 may rightly protest that this is a level of semantics, of meaning, which they never intended to solve and this is a reasonable position. Nonetheless, the problem remains. Without a means of separating these different kinds of meanings, we shall not have a semantic web, which can address the complexities of culture. Indeed, we need more, because these meanings also change historically, such that a term, which meant one thing in the 17th century may mean something very different today. Hence the word nice, which in the 17th century frequently meant lazy or lewd, or lascivicious, now means something quite different when persons speak of a nice day. We need new kinds of search engines which do not simply search for a "natural language" term, but allow us to distinguish between local, regional, national, and international levels, multi-lingually, multi-culturally and historically (i.e. including etymologies).

4. Computer Science and AI

Within the field of computer science and particularly among a small group of individuals in Artificial Intelligence (AI), semantics has a much narrower meaning. Here the quest is to arrive at a supposedly objective machine-readable code whereby machines can make decisions without human intervention. In this context, meaning is reduced to efficient commands and decision trees. There is an assumption that if the code were perfected then humans would no longer be necessary. For instance, computer scientists such as Carl Hewitt have claimed that one needs to replace humans with robots in the case of decision systems. The quest is to create machines:

> that could take care of us, that could be our guardians and that would also be our rulers and policemen…to program computers and robots that could garner all the weapons of mass destruction into a machine controlled system, in the same way that you have to take matches away from children.[9]

According to the supporters of this school, all decision making concerning military actions, when to send planes, throw bombs etc. needs to be removed from the human

sphere and the goal is to turn the keys[10] for all such actions to robots. To this end the army, navy and the air force are all working on autonomous decision robots[11:]

> The necessary turnover in personnel you get in human-based systems, because of their very short lifetimes, seems to throw instability into the system. And the general diversity of human stock we have, in terms of different languages, cultures and interest is not something that can be smoothed out very quickly.[12]

In this approach the subjective meanings of humans with their many languages, cultures and attendant ambiguities are merely a nuisance and ultimately meaningless. The profound dangers of such a quest were pointed out nearly three decades ago by the Nobel physicist, Joseph Weizenbaum (1976):

> The computer has thus begun to be an instrument for the destruction of history…. For when society legitimates only those 'data' that are 'in one format' and that 'can easily be told to the machine' then history, memory itself, is annihilated…. And the curious paradox is that the immortality of knowledge means the death of culture.[13]

These dangers were restated a decade later in Grant Fjermedal's *The Tomorrow Makers* (1986), a fascinating book on the development of living brain machines.[14] Fjermedal noted that this vision of autonomous decision robots was a quest for a non-biological intelligence which, according to Richard Jarrow, founder of NASA's Goddard Institute, was destined to replace humans altogether.[15]

This goal of creating autonomous decision robots helps to explain a growing fascination with and commitment to natural language and so-called common sense worlds, which were described by Jerry Hobbs and Robert Moore (1986).[16] It helps explain also the rise of artificial intelligence projects such as Doug Lenat's CYC, Generic Artificial Consciousness (GAC) and Common Sense.[17] It suggests a deeper reason for the Defense Advanced Research Projects Agency's (DARPA) very active participation in Knowledge Query Markup Language (KQML), Knowledge Interchange Format (KIF), DARPA Agent Modeling Language (DAML) and, possibly, their increasing role in W3's quest for a semantic web.

One is tempted to dismiss such a quest to replace human intelligence by machines as efforts of a marginal minority in the military. However, analogous ideas are being developed in the realm of American industry. For instance the authors of Visionary Manufacturing Challenges for 2020 foresee new techniques evolving independently of language and culture, which is the opposite of the European approach:

> A major task will be to create tools independent of language and culture that can be instantly used by anyone, regardless of location or national origin.
> Tools will have to be developed that allow for effective remote interaction. Collaboration technologies will require models of the dynamics of human

> interactions that can simulate behaviors, characteristics, and appearances to simulate physical presence.[18]

By implication there are two fundamentally different visions of a semantic web. One aims at understanding human meanings, which vary from place to place and vary historically. A second aims to use natural language and common sense to offer a single language for robots acting independently of humans with no reference to cultural diversity and the complexities of history. In our view, the first vision needs to be developed. The second is misleading and dangerous. It implicitly undermines the larger vision of the W3 Consortium as a world wide web for humans. Ultimately the second vision is a threat to the human race.

5. Globalism

Historically, these have been other, more subtle, trends working against multilingualism. Ever since the scientific revolution in the Renaissance there has been a gradual tendency towards international standards which gained enormous ground in the nineteenth and twentieth centuries with the rise of many international organizations such as the International Standards Organization (ISO), International Telecommunications Union (ITU), and the United Nations Educational Scientific and Cultural Organization (UNESCO). Underlying these bodies was a vision that one needed to reach agreement on terms in order to make progress. Local and regional agreement were first steps, national agreement was one step further and international agreement on a term or concept was ultimately the goal.

In the realms of science and technology this is essential. Science is concerned with universally valid laws/rules. Hence we need globally accepted definitions of zinc, chemical formulae and the like if we are to have an international scientific community. This is also the case in medicine. Our definition of a heart needs to be the same if surgeons are to operate successfully around the world. This quest also relates to Tim Berners Lee's assumption that meaning is closely linked with logic and thus with things which can be proven. Hence his notion of a semantic web strives for information/ knowledge that is universally true and the same everywhere.

In the realms of the arts and culture, however, the situation different for three fundamental reasons. First, the cultural sector has a historical dimension, which is central to its existence. In the case of science, the focus is on the laws/rules, which apply now.19 In culture, the arts and the humanities, the historical commentaries on great authors such as Homer and Shakespeare or on great artists such as Leonardo and Rembrandt are not just of passing interest. They are central to the field, for the depth of culture lies precisely in the cumulative effect of these historical commentaries over the ages. Indeed these commentaries over time give cultural objects such as the text of Shakespeare's *Hamlet* their full importance. Hence, whereas science deals with laws, rules, formulae, which function as if they were a-temporal, cultural objects entail an essential temporal dimension. In science, a database of current formulae and definitions may be sufficient.

In the realm of culture we need databases, which include historical definitions, (etymologies) and make visible the cumulative dimension of cultural objects.

Related to this is a second difference. The goal of science is to arrive at truths or at least working hypotheses concerning which there is global acceptance. The greater the acceptance the more scientific a claim becomes. In the cultural sector, global agreement is extremely rare. Even in the case of UNESCO World Heritage sites there is often disagreement about what should be included. Indeed the richness of the cultural sector lies precisely in the amount of disagreement; in the diversity of interpretations concerning the same object. Hence, whereas science needs databases to record those "facts" on which there is global agreement, culture requires databases to record all the disagreements concerning a given cultural object.

Hence the semantic web as it is emerging reflects admirably the needs of modern science and technology. But it does not yet answer the more complex needs of the cultural sector. Some might argue that this is not essential and merely a luxury. In a world where narrow identities of fundamentalist sects are threatening the very fabric of society the need for identities with dimensions of tolerance many become our only hope for long-term survival as a civilization. Meanwhile, economists who wish to insist only on financial dimensions, need reminding that culture is intimately connected with tourism, which is the most important source of income in all the G7 countries and many other countries of the world. In addition to being fundamental to our sense of identity, it is thus also one of our most important sources of economic gain.

6. Cultural Semantics

There is a third reason why culture is different from science and technology. Science is concerned only with the globally accepted laws/rules. Cultural objects/ products have local, regional and national variants. To take a prosaic example: beer has certain international standards, which are necessary to assure that the brew is safe and not poisonous. But ultimately what makes beer interesting is that German beer is different from Dutch or Danish beer. Even within a region and locally there are many variants.

To take a more exalted example: paintings of the *Annunciation* are culturally rich precisely because there are so many national, regional and local variants. Hence a semantic web, which aims to create databases with only a single definition of beer or of only one *Annunciation* is not useful. In the case of cultural products/objects we need databases to indicate information/ knowledge at the global, international, national, regional and local levels. And in an increasingly networked world we need evermore links between these levels.

Given the global nature of science, it is ultimately sufficient that there is only a single term for a given law, principle, rule or concept in a single language. Nuclear physics or radio astronomy do not preclude multilingualism, but one could argue that multiple languages only risk adding further confusion to an already complex subject. By contrast, in the cultural sector local, regional and national variants are essential to the richness of cultural expression, and depend fundamentally on different languages and dialects. Thus

a semantic web, which includes cultural, spatial (local, regional, national, global), historical and interpretative dimensions is one of the essential challenges that faces us in the future.

Since the rise of the nation state there has been a tendency to compartmentalize knowledge. Local knowledge was stored locally, regional knowledge at the provincial or state level, national knowledge in the capitals of countries and international knowledge was stored in a few global libraries such as the Vatican and more recently in national collections (e.g. Bibliothèque Nationale de la France, Library of Congress).

The advent of new technologies and the Internet led in a first instance to a networking of the great international libraries and research institutions such as the Research Libraries Information Network (RLIN) and through projects such as the Gateway to European National Libraries (GABRIEL). Such networks provide access to tens of millions and potentially hundreds of millions of titles. Through projects such as Gallica (BNF, Paris) the full contents of such titles are also becoming available.

Meanwhile, our search engines often implicitly assume that everything on the web is equally valid. Alternatively they perpetuate nineteenth century, positivist assumptions about terms: i.e. that, implicitly, when we search for a word a single definition is entailed. The quest to achieve interoperability of content further strengthens this trend. There is an assumption that unless there is complete equivalence between the meanings of fields, there can be no interoperability. Paradoxically, however, if there is a complete equivalence in contents of fields there is nothing gained in bridging meanings at different levels. Complete interoperability in this narrow sense would lead to precisely the McWorld effect against which Barber warned.[20]

Needed therefore is a more subtle approach. We need more than just the internationally agreed upon usage of a term. We need access to national, regional and local versions, with an indication at each stage about the level of agreement that exists concerning a term in a given language or dialect. Hence, when we search for heart the system needs to provide us with terminology and a definition, which have been agreed upon internationally and at the same time indicate national, regional and local variants. If the local interests us there may be cases where a local term is a) defined in a local dictionary or dialect phrasebook; b) where it is available in a recorded corpus and not yet formally defined or c) where it is used locally and not yet even systematically recorded. Until we have a framework, which allows such distinctions we cannot achieve full syntactic and semantic interoperability. Hence a challenge lies in a new synthesis of knowledge at local, regional national and international levels complete with new methods for reflecting these levels within our search engines and devices for navigating through networked knowledge. This is the challenge of cultural semantics.

7. Conclusions

The first half of the twentieth century introduced new ideas for computers, which transformed earlier concepts of computational devices which have evolved since the

times of Pascal and Leibniz. The last half of the twentieth century transformed the notion of individual computers to an inter-networked world, whereby supercomputers and personal computers can be linked through computational grids. The notion of computers as devices concerned only with computation, number crunching, evolved also to include text, images, sound, touch and more recently smell and taste.

The 21st century marks a new epoch in these developments. In 1995 there were 30 million users. In 2000 there were 300 million users and in the past two years the Internet the has grown to over 544 million users. This figure is predicted to double in turn within the next five years. Within a decade more persons will have access to the Internet than has ever been the case with any other technology.

Freud, McLuhan, Levy, and others have argued that computers should be seen as extensions of man: not only in the physical sense of mechanical tools, but also in a conceptual sense. Kurzweil would go further to claim that computers are extensions of man in a spiritual sense. In this context, the vision of a semantic web is one of the keys to the future. We need to get beyond number crunching and word crunching in order to get at the meaning of texts, images, and other creations of the human spirit.

We have noted that there are at least four approaches to the semantic web:

> 1) The W3 Consortium led by the vision of Tim Berners- Lee focusses on semantics in terms of logic.
>
> 2) The Dublin Core (Metadata Initiative) limits semantics mainly to the meaning of metadata elements/fields rather than the contents of those elements/fields.
>
> 3) A small group within the AI community sees semantics strictly in terms of machine-readable instructions, which permit autonomous software agents and hardware robots to operate and make decisions in the absence of humans.
>
> 4) Cultural semantics entails a commitment to meaning, which takes into account multi-lingual, multi-cultural, and historical dimensions at the local, regional, national and international levels.

We have suggested that the efforts of 1) the W3 Consortium thus far are important, very useful for transactions, but do not yet answer the needs of human meaning; that the efforts of 2) the Dublin Core mark another important step forward, but that this cannot be seen as a comprehensive solution. We suggested that the approach of 3) a small minority in the AI community potentially undermines the vision of the W3 and is ultimately a threat to the human condition. What we need is a semantic web, which embraces cultural dimensions, which provides new levels of access to knowledge at the local, regional, national as well as the international levels. The essence of science may lie in the universality of its claims, in universals. The essence of culture lies in the unique, in particulars, in the exceptions to the rule. We have exceptional databases for the universal laws of science but we have very little by way of databases for the unique and

exceptional expressions of culture. To achieve this is one of the great challenges for the semantic web of the future: not to replace humans, but rather to find new ways of making visible their abiding expressions.

Maastricht 23 02 2002

8. References

[1] For a longer discussion of this theme see the author's: "Syntactic and Semantic Interoperability, New Approaches to Knowledge and the Semantic Web," *New Review of Information Networking*, Springer Verlag, Berlin, 2002, 16 pp. (Volume 7) (in press) and Understanding New Media: Augmented Knowledge and Culture.630pp. (in press).
[2] Cf. the book by Viktor F. Frankl, *Man's Search for Meaning: an introduction to logotherapy*, translated by Ilse Lasch, Beacon Press, Boston, 1962.
[3] Ernst Cassirer, *Substanzbegriff und Funktionsbegriff, Untersuchungen über den Grundfragen der Erkenntniskritik*, Bruno Cassirer, Berlin, 1910. English translation: Substance and Function, Open Court, Chicago, 1923. These ideas were developed in his *Philosophie der symbolischen Formen, Bd. 3: Phenomenologie der Erkenntnis*, B. Cassirer, Berlin, 1923-29. English Translation: Philosophy of Symbolic Forms, Volume 3: *Phenomenology of Knowledge*, Yale University Press, New Haven, 1957. These ideas were further popularized in Cassirer's, *The individual and the cosmos in Renaissance philosophy*, Barnes and Noble, New York, 1963.
[4] http://www.w3.org/2001/sw/.
[5] A slightly different arrangement is given by Rohit Khare, XML. "The Least you need to Know."
See: http://www.cs.caltech.edu/~adam/papers/xml/tutorial:

Syntax	SGML
Style	CSS/XSL
Structure	HTML
Semantics	XML

[6] This is a subset of Standard Generalized Markup Language (SGML).
[7] Cf. John Sowa, "Ontology, Metadata, and Semiotics," *International Conference on Conceptual Structures, ICCS'2000, 14-18 August 2000*, Darmstadt, Germany.
See: http://www.bestweb.net/~sowa/peirce/ontometa.htm.
The distinction between syntax, semantics and pragmatics comes from Peirce who saw these as the three branches of semiotics. Charles Sanders Peirce, "On the algebra of logic," *American Journal of Mathematics*, vol. 7, 1885, 180-202; *Collected Papers of C. S. Peirce*, ed. by C. Hartshorne, P. Weiss, & A. Burks, 8 vols., Harvard University Press, Cambridge Mass., 1931-1958. Particularly vol. 2, 229.
[8] See: http://dublincore.org/about/history/
[9] Grant Fjermedal, *The Tomorrow Makers, A Brave New World of Living Brain Machines*, Tempus Books, Redmond 1986, 141.
[10] Ibid., 144. Asked what would make persons take this step the answer was fear caused by "small nuclear wars popping off here and there- like between India and

Pakistan, or between Israel and the Arabs." In the post September 11, 2001 world these claims of 1986 seem frighteningly prescient.
[11] Ibid., p. 121
[12] Ibid., p. 143.
[13] Joseph Weizenbaum, *Computer Power and Human Reason. From Judgement to Calculation*, W. H Freeman and Co., New York, 1976 (Published Penguin/Pelican Books, 1984, 238.
[14] Grant Fjermedal, *The Tomorrow Makers* (1986) as in note 9.
[15] Ibid., 139.
[16] Jerry R. Hobbs, Robert C. Moore, *Formal Theories of the Commonsense World*, Norwood, Ablex Publishers, Norwood, NJ, 1985 (Ablex Series in Artificial Intelligence, Vol 1).
[17] "Battle of the Brains", *Wired*, November 2001.
[18] *Visionary Manufacturing Challenges for 2020*, ed. Committee on Visionary Manufacturing Challenges, Board on Manufacturing and Engineering Design; Commission on Engineering and Technical Systems; National Research Council Washington: National Academy Press, 1998.
See: http://bob.nap.edu/readingroom/books/visionary/ch2.html#gc3
[19] To be sure there are historians of science who remind us that the history of the subject is useful in understanding how we got to where we are today, but this is seen more as luxury than as an essential prerequisite for the advancement of science.
[20] Benjamin R. Barber, *Jihad vs. McWorld*, Times Books, New York, 1995.