The "Emergent" Semantic Web: A Consensus Approach for Deriving Semantic Knowledge on the Web[1]

Clifford Behrens and Vipul Kashyap
Telcordia Technologies, Inc
445 South Street
Morristown, NJ 07960-6438, USA
{cliff, kashyap}@research.telcordia.com

Abstract. The recent and growing interest in the Semantic Web has given rise to a flurry of activity in standardization bodies (such as the W3C) to specify semantics using formal languages and inference mechanisms. The real challenge, however, is to link formal semantics with deeper meaning as reflected by consensus discovered among users on the Semantic Web. We believe the process of deriving and formally describing ontologies for the Web (expressed using standardized languages) is necessarily a social-cultural one; hence, new consensus-based tools are required to derive shared semantic systems for different communities of interest. This paper introduces Consensus Analysis as a means for deriving semantic knowledge from the information provided by subject matter experts and describes the *Schemer System* prototype for acquiring and processing this information. The results of a trial application of this approach and prototype on technologists asked to identify current mass market consumer trends in the domain of Internet privacy and security are reported. These findings implicate Consensus Analysis as a powerful tool capable of enabling the semantic Web by yielding core knowledge such as controlled vocabularies and domain ontologies.

1. Introduction

Recently there has been a great interest in the Semantic Web and issues related to specification and exploitation of semantics on the WWW. In particular, shared ontologies are being proposed for representing the core knowledge that forms the foundation for semantic information on the Web. Fensel [1] has identified two broad research thrusts related to ontologies:

1) Approaches to standardize the formal semantics of information to enable machine processing. Work being done as a part of the W3C RDF working group [2] and the DAML+OIL initiative [3] falls within this category.
2) Approaches to define real world semantics linking machine processable content with meaning for humans based on consensual terminologies.

To realize the goals of the Semantic Web, there is a need to wed approaches centered on the formal representation of semantics with approaches to systematically acquire terminologies that best express shared systems of meaning among users. We believe the process of deriving and describing domain ontologies necessarily involves the search for consensus among domain experts and, therefore, is inherently a social-cultural one. As such, the proper approach to deriving knowledge, like domain ontologies, ought to be sensitive to the semantic context of information, and should be informed by the real-world bottom-up, decentralized process in which knowledge typically evolves. After all, decentralized models for consensus achievement better reflect the dynamic sociological characteristics of the Web (which have been the cause for its rapid acceptance and success). In this manner more meaningful ontologies (expressed in standardized formal languages) can emerge through more natural interactions of Web users within their respective communities. We agree with Fensel [1] who claims that the real challenge for making the semantic Web a reality is, "a model for driving the network that maintains the process of evolving ontologies."

Consistent with this claim, similar interest in Knowledge Management processes has motivated new research in automatic knowledge acquisition, classification, and representation [4]. Much of the discussion on Knowledge Management has focused on information technology, e.g., hardware, software and communications networks, but has not laid out in clear terms what notions of "knowledge" need to be supported by this technology. For example, there exists in the literature a recurring theme that knowledge is any information stored in a Knowledge Repository, and that this knowledge can somehow be acquired or "discovered" automatically from disparate, heterogeneous information sources, e.g., Web pages and networked document collections. This approach seems at best naïve as it ignores the context and intended purpose of source information. Without establishing this context and purpose, it seems unlikely that much useful "knowledge" can be discovered as it leaves matters pertaining to information's meaning and relationships with existing knowledge open to broad interpretation.

Within the literature there is expressed the idea that not all information is knowledge; information only becomes knowledge once it is mapped to a knowledge structure, i.e., it is organized in a way that makes it accessible and comprehensible to users [4]. In fact, this qualification suggests that there may exist many such structures for organizing the same information, again supporting the idea that context and purpose are essential for transforming information to knowledge. It also implicates the importance of knowing the "community of interest" (COI) for both the producer and consumer of information to enable this transformation since members of different COIs may set different contexts or use the information with very different intentions. In the emergent Semantic Web, it is critical to determine the “consensus” knowledge structures for a COI.

The term "community" is becoming ubiquitous, particularly in discussions related to delivery of personalized services on the Internet, yet there exist distinct usages of the term. For some, a community seems to consist of all who, because of a shared interest in certain kinds of information, frequent the same place, real or virtual, regardless of any interaction among them, e.g., all those who browse the same Web site [5]. A more social usage entails information exchanges among a collection of individuals, e.g., all those who exchange useful information about some topic of mutual interest through email or chat rooms [6]. A more sophisticated "cultural" notion of community refers to all who, in addition to meeting the two preceding conditions, share a vocabulary, *semantics* and theory for organizing information.

For members of this class of community there exists some common purpose and key concepts for communicating ideas and sharing experiences. In this paper, this last notion of community is adopted because, as it will be demonstrated, it provides an opportunity to analyze knowledge, and its variations among individuals, with greater formal rigor. It also helps to more clearly draw the lines operationally between information, individual knowledge, and what will be referred to later as "cultural knowledge."

Much research has already been conducted in the social sciences, particularly cognitive anthropology and cognitive psychology, on modeling knowledge domains, i.e., conceptual categories that include other semantically-related categories, and eliciting the information needed to build these models. However, most of these methods are extremely time-consuming, taxing the attention of a few SMEs (Subject Matter Experts), those recognized as experienced and possessing specialized domain knowledge. *Consensus building* is another approach to building knowledge representations that is gaining increasing popularity in the Information Processing standards community and elsewhere [7, 8]. New information technology could be applied to eliminate much of the need (and enormous cost) of face-to-face group decision-making meetings, e.g., read [9] and [10] for examples of IT approaches to collaborative knowledge construction.

Previous methods for building knowledge from consensus have been tried, e.g., Delphi approaches [11, 12], but these are typically iterative and require much human intervention. While the importance of consensus to achieving views that best represent collective thinking is often stressed, too often views are biased strongly by the force of individual personalities and are not representative of any particular COI. Other problems arise from the heterogeneous composition of decision-making groups whose members conceptualize the same problem from widely different perspectives, i.e., those of different COIs. Moreover, simple polling methods that only average expert opinion do not usually yield results with the depth and logical properties of real domain knowledge, nor do they exploit the contributions of the most competent SMEs. Thus, there is need for a different approach that derives, rather than forces, consensus views, does so without the need for much human intervention and many iterations, acquires useful information from SMEs (weighted by their competence) at their convenience, and is capable of yielding shared knowledge for a demonstrable domain of interest.

By combining new formal and more rigorous approaches to consensus-modeling, specifically powerful methods of Consensus Analysis that already have been tested successfully by Cognitive Anthropologists in numerous knowledge domains, the network services approach taken in this research overcomes the limitations of previous computer-assisted approaches. This is accomplished by (1) incrementally refining or "evolving" knowledge, (2) providing *metrics* for evaluating the cultural saliency of a domain and the knowledge-based competency of SMEs in a COI, (3) dynamically assigning SMEs to the most “appropriate” COI and (4) not only spreading the task of knowledge acquisition among many SMEs, rather than just a few, but also leveraging Web infrastructure to engage them at their convenience.

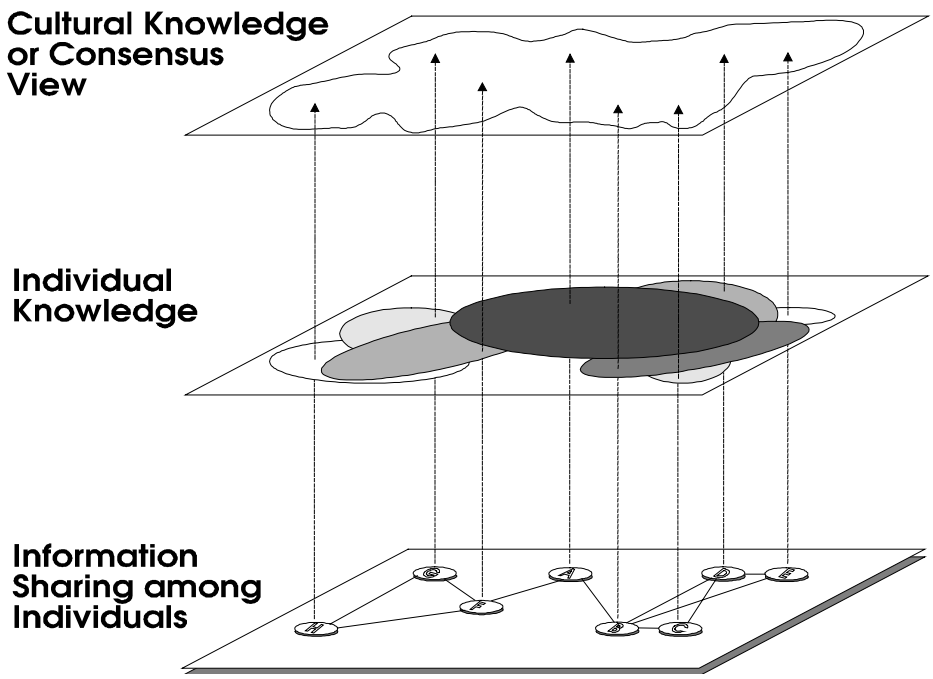

Figure 1. Knowledge distribution, knowledge sharing and consensus.

2. Cultural Knowledge and Consensus Analysis

Consensus Analysis is based on a few simple, but powerful, ideas , i.e., *knowledge is both distributed and shared* [13]. For any knowledge domain, and any group of subjects "expert" in this domain, so-called "SMEs" possess different experiences; hence, they know different things, and some of them know more than others (see Fig. 1). Information sharing, e.g., among individuals A-H in the figure, facilitates the availability of a much larger pool of information with non-uniform distribution of knowledge across SMEs. For example, many information standards groups are composed of data providers, data users, librarians and software vendors. These groups tend to possess different experiences with data, and apply their own unique views and semantics to describe these data. Yet, certain individuals (the hi-tech "gurus") are recognized as being more knowledgeable than others, i.e., there exist recognized domain "experts." Because of their widely regarded and highly-valued knowledge, these experts are frequently requested to share what they know with others as consultants or as leaders in standards-setting groups, or render opinions about how best to describe or classify information in their domain of expertise. Hence, one typically finds within any COI that there is differential expertise among its members, but also some knowledge that is widely-shared and recognized as being "essential." In fact, this knowledge may be so fundamental and its use so widespread that, over time, it becomes logically well-structured or canonical, e.g., even published as a metadata content standard. *The process of mapping information onto such a consensus standard is the essence of cultural knowledge creation.*

Cultural knowledge is not all that one knows (e.g., the set of knowledge for each individual represented in the middle layer of Figure 1); nor is it the sum total of what

everybody knows (e.g., the union of individual knowledge sets in the middle layer). Rather, it is an abstraction, knowledge shared in its "broad design and deeper principles" by members of a society or community [14]. In other words, while its entire details are not usually known (or can be always be articulated explicitly) by anyone, cultural knowledge consists of those things that all members of a COI understand all others hold to be true.

Kroeber [15] referred to this highly-structured, rich form of knowledge as a "systemic culture pattern," a coherent subsystem of knowledge that tends to persist as a unit. This unit features a semantic system, consisting of an appropriate vocabulary and grammar, for classifying and talking about elements within a knowledge domain. Examples of cultural knowledge are: a kinship terminology [16, 17], or perhaps a metadata content standard [18], a consensus statement for screening cancer [19], or a set of software requirements [20]. It is this shared pool of structured information, acquired primarily by learning, which constitutes cultural knowledge [21].

2.1 Robustness of the Consensus Model

The significance of information sharing and distribution of cultural knowledge has encouraged some researchers to exploit consensus, measured by intersubject agreement, as an indicator of knowledge. The method of Consensus Analysis was first presented in several seminal papers [13, 22, 23]. In addition to introducing the formal foundation for Consensus Analysis (reviewed later in more detail), the initial papers cited above also provided examples of its application to modeling knowledge of general information among US college students, and the classification of illness concepts among urban Guatemalans. Other more recent applications of Consensus Analysis have focussed on measuring cultural diversity within organizations [24]. These successes, obtained for a wide variety of domains and social-cultural contexts, indicate that the following three explicit assumptions, upon which Consensus Analysis is based, are extremely robust [13]:

i) *Common Truth*. There is core knowledge (expressed in a highly probable set of answers to questions or "items") that is "applicable" to all SMEs or, put another way, all SMEs are members of the same COI and generally share a common perspective or "cultural reality."

ii) *Local Independence*. The information or responses provided by each SME are acquired independently from those of other SMEs, i.e., SME item response random variables satisfy conditional independence for all possible response profiles and the core answer set.

iii) *Homogeneity of Items*. Each SME has a fixed level of "competence" or "expertise" across all items, i.e., items used to sample what SME's know are equally difficult and provide representative coverage of a coherent domain. In practice, this assumption has been found to be quite robust and requires only that those SMEs who are most knowledgeable in a domain consistently outperform non-experts.

From these assumptions, it is possible to derive a method for estimating three properties of interest: (1) a measure of the overall saliency of the knowledge domain represented by the pool of items, (2) the level of domain expertise or "cultural competence" for each SME based on the amount of consensus or agreement between his/her responses to items with those of all other SMEs, and (3) the most probable set of "correct answers," inferred from the responses

of each SME and weighted by their respective competence measures, i.e., the consensus view.

2.2 Statistical Methodology

As mentioned earlier, the Consensus Analysis Model can be derived formally from the three assumptions given in section 2.1. This formal model consists of a data matrix X containing the responses X_{ik} of SMEs *1..i..N* on items *1..k..M*. From this matrix another matrix M^* is estimated and it holds the empirical point estimates M^*_{ij}, the proportion of matching responses on all items between SMEs *i* and *j*, *corrected for guessing* (if appropriate), on off-diagonal elements (with $M^*_{ij} = M^*_{ji}$ for all pairs of SMEs *i* and *j*). Alternatively, another matrix C^*, which contains the observed covariances C^*_{ij} between the responses of SMEs *i* and *j*, corrected for variance among SME answers, may be substituted for M^* [25]. To obtain D^*_i , an estimate of the proportion of answers SME *i* "actually" knows and the main diagonal entries of M^* (or C^*), a solution to the following system of equations is sought:

$$M^* = D^* D^{*\prime} \text{ or alternatively,} \quad (1)$$

$$C^* = D^* D^{*\prime} \quad (2)$$

where D^* is a column vector containing estimates of individual competencies $D_1..D_i..D_N$ and $D^{*\prime}$ is merely its transpose. Since equation 1 (or 2) represents an overspecified set of equations and because of sampling variability, an exact solution is unlikely. However, an approximate solution yielding estimates of the individual SME competencies (the D^*_i) can be obtained by applying *Minimum Residual Factor Analysis* [26], a least squares approach, to fit equation 1 (or 2) and solve for the main diagonal values. The relative magnitude of eigenvalues (the first eigenvalue λ_1 at least three times greater than the second) is used to determine whether a single factor solution was extracted. All values of the first eigenvector, v_1, should also range between 0 and 1. These results test the validity of the Common Truth assumption.

If the criteria above are satisfied, then the individual SME competencies can be estimated with:

$$D^*_i = v_{1i} \sqrt{\{\lambda_1\}} \quad (3)$$

The D^*_i, then, are the loadings for all SMEs on the first factor. These estimates are required to complete the analysis, i.e., to infer the "best" answers to the items. The estimated competency values (D^*_i) and the profile of responses for item *k* ($X_{ik,l}$) are used to compute the Bayesian *a posteriori* probabilities for each possible answer. The formula for the probability that an answer is "correct" follows:

$$Pr(\langle X_{ik} \rangle\ i=1 \mid Z_k=l) = \prod_{i=1}^{N} [D^*_i + (1-D^*_i)/L]^{Xik,l}\ [(1-D^*_i)(L-1)/L]^{1-Xik,l} \quad (4)$$

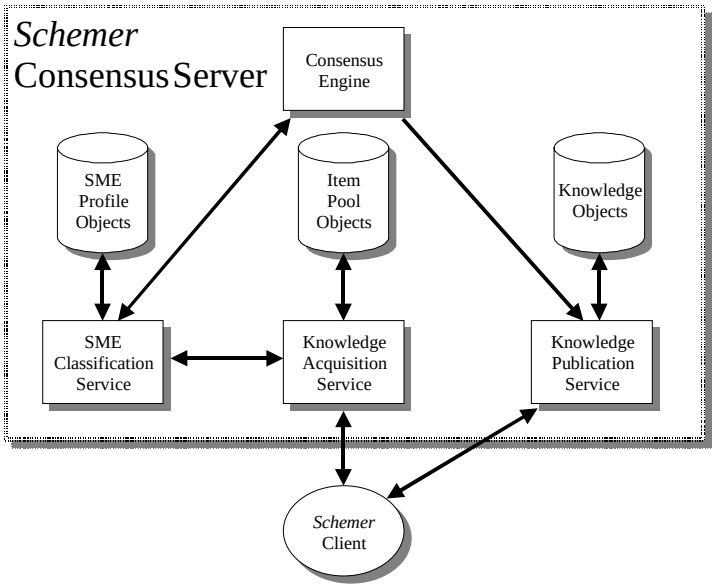

Figure 2. *Schemer* system architecture.

where Z_k is the "correct" answer to item k, l is the l^{th} response to item k, and L is the total number of possible responses ($l_1 ... l_L$) to item k. Again, it should be mentioned that the "correctness" of an answer is relative to the perspective shared by members of a particular COI, i.e., the one sampled. Equations 1-4 provide formal motivation for the approach taken in this research, and indicate algorithms that need to be implemented in software as part of a network-enabled consensus server.

3. System Architecture and Prototype

Telcordia researchers have begun to design a software prototype called the *Schemer System*, shown in Fig. 2. Key software components in this design have already been implemented to communicate better some of the objectives of the approach, stimulate greater interest in it, and demonstrate the feasibility of automating Knowledge Acquisition and Consensus Analysis modeling. Future work will include development of Publication Services and a fuller integration of software components in a continuous Web-based service.

In our current design, the *Schemer System* consists of a *Schemer Client* and *Schemer Server*; however the latter really involves the interaction of four services: a *Subject Matter Expert Classification Service*, a *Knowledge Acquisition Service*, a *Consensus Engine*, and a *Knowledge Publication Service*. These services read and write information to several data bases, one storing information about SMEs, another storing pools of items used to acquire information from SMEs, and another which stores the derived knowledge structures, i.e., the controlled vocabularies, forecasts, ontologies, classification schemes, or productions systems. Next, each of these services is examined in more detail.

Client. The job of the *Schemer Client* is to provide a graphical/text interface through which a user communicates with the *Schemer Server*. It presents information sent by the Server such as item forms and knowledge visualizations, both textual and graphic.

SME Classification Service. This service determines a knowledge domain of interest for a SME, and assigns a SME to his/her proper COI. Classification is necessary to present a SME with meaningful items and knowledge derived from a consensus analysis of peer responses. Knowledge domain identification may be determined either by asking a SME to select a known knowledge domain from a list or, if unknown to *Schemer*, the SME is asked to input the name of this knowledge domain. COI classification may be accomplished in two stages, *a priori* and *post hoc*. If nothing is known about the SME, then preliminary COI classification is made by asking the SME to choose a COI from a list of COIs already known to *Schemer*. If, however, the SME cannot find an appropriate COI in this list (or if the knowledge domain is unknown to *Schemer*), then he/she is prompted for a list of key terms frequently used to describe objects in the knowledge domain of interest. These terms are matched against key term lists (if they exist) for known COIs to determine the "best" COI classification for this SME. But once an item form has been constructed for this SME and used to acquire more information, *post hoc* analysis of results obtained from Consensus Analysis may be used to reclassify this SME, if he/she so chooses. To compare new information obtained from the SME with information known for SMEs already classified, the *Subject Matter Expert Classification Service* reads the data it needs from a repository of SME profile objects.

Knowledge Acquisition Service. Based on the knowledge domain and preliminary COI classification obtained for a SME, this service selects an appropriate item pool object and composes an instrument or form that is used to determine what the SME knows about the domain. Items published on these forms are read from a repository of item pool objects, each identified by knowledge domain and COI. This service sends the form to the Client where the SME enters his/her responses to items on the form, then sends these back to the *Knowledge Acquisition Service*. The SME's response pattern, along with his/her ID, is stored in a repository of SME profile objects, also grouped by COI and knowledge domain.

Consensus Engine. This service performs a Consensus Analysis of data collected for a knowledge domain/COI grouping each time new responses are added to a SME's profile, and stores the updated result in a *Knowledge Repository*, along with ancillary statistics, e.g., Goodness-of-fit indices. Not only does the *Consensus Engine* analyze data read from SME profiles, but it also adds information to these, e.g., a SME's competency score.

Knowledge Publication Service. On a user's request, this service constructs forms with textual and graphical representations of derived knowledge, stored in the *Knowledge Repository*, for presentation on the *Schemer Client*. Access to information stored in this repository is also provided by this service so that a user can retrieve a knowledge object for use with his/her own software application.

To date, a skeleton *Knowledge Acquisition Service* has been built, capable of taking as input from a SME's Web browser a knowledge domain and COI value, then return a form with an appropriate item set for this knowledge domain/COI combination. Currently, only dichotomous (True/False) formats are supported. Once a SME completes this form and submits his/her responses to the *Knowledge Acquisition Service*, it notifies the *Consensus Engine* that a SME's profile has been updated. The *Consensus Engine* processes all of the response vectors for SMEs in the same knowledge domain/COI data base, then stores the

results, e.g., eigenvalues, SME competency scores and the estimated answer key, in the knowledge base. All of these services have been implemented in Java® and the R® statistical programming environment, so can run under Unix® or Windows®.

4. Experiment

The remainder of this paper describes an experiment that was conducted among Telcordia technologists to derive a consensus view of mass-market consumer trends related to Internet security and privacy. While no attempt will be made to derive a domain ontology from this experiment, our intent is to demonstrate how Consensus Analysis works and to further suggest that it seems well-suited for this purpose.

A prototype of the *Schemer* system was built and used to deliver a questionnaire consisting of sixty-seven items related to privacy and security of information on the Internet (see Appendix A). These items were derived from Georgia Tech's *10th World Wide Web User Survey*, which includes a section entitled, "Online Privacy and Security". A request was mailed electronically to Research Scientists belonging to two labs within Telcordia Technologies Applied Research. These sample SMEs were asked to answer items on the questionnaire *as if they were domain experts* being asked for their opinions about *mass market consumer trends within the Web user community*, not necessarily with their personal opinion. Along with responses to the questionnaire, SMEs were asked for their employee ID and a list of no more than twenty descriptors that they believed best represented their professional area of expertise. The former was used as a pointer to other ancillary information about the SME, e.g., lab, department, group, and office location, while the latter was collected to help associate the domain expertise of a SME with that of others in the sample. A total of thirty-six Research Scientists responded to the request above. This sample was opportunistic, not random; moreover, a special request was made to members of Telcordia's *Computer Networking Research* department, which specializes in Internet security issues, so that the responses of these domain experts could be compared to others in the sample.

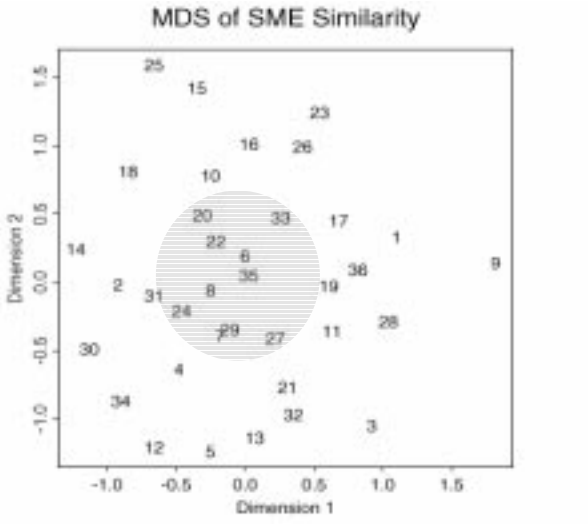

Figure 3. Plot of MDS results showing similarities in responses among SMEs. Similar SMEs are plotted close to one another. Stress= 0.260 after 19 iterations.

The similarity or agreement among SME response patterns can be explored in Figure 3. This two-dimensional plot was obtained from a Multidimensional Scaling of only the off-diagonal entries of the consensus matrix (M_{ij}* in Equation 1) calculated for the thirty-six Telcordia SMEs [27]. In this visualization, the SMEs with similar responses are plotted closest to one another. The absence of clustering in this plot suggests that the study sample was drawn from a single COI whose members share core domain knowledge about "Online Privacy and Security." This notion was tested more rigorously by estimating a consensus model for these data.

4.1 Knowledge Domain Validation

As the review of Consensus Analysis pointed out, knowledge derivation rests on establishing the validity of the domain to those SMEs in the sample. This is accomplished by inspecting the relative magnitudes of the eigenvalues for the first factors extracted from the consensus matrix using Minimal Residuals Factor Analysis. Again, the "rule-of-thumb" is that the eigenvalue of the first factor must be at least three times greater than the second; moreover, subsequent eigenvalues should all be small and roughly equivalent. Inspection of the eigenvalues for the first three factors extracted from the response set collected from Telcordia SMEs reveals that the first is over six times greater than the second, and the second and third eigenvalues are almost equal (see Table 1). This lends strong support to the claim that the items on the questionnaire are sampling a single, coherent knowledge domain, and that this domain has salience for the sample of respondents. Moreover, the high Pseudo-Reliability Coefficient (0.944) also obtained suggests that these results are stable and would likely be the same ones obtained with repeated sampling [13].

Table 1. Eigenvalues for testing saliency of "Online Privacy and Security" knowledge domain.

Factor	Eigenvalue	Percent	Cumulative %	Ratio
1	11.902	77.8	77.8	6.500
2	1.831	12.0	89.8	1.175
3	1.559	10.2	100.0	
	15.292	100.0		

4.2 Estimation of SME Competence

Having established the saliency of "Online Privacy and Security" as a knowledge domain for SMEs in the sample, it is possible to estimate each one's competency in this domain. The competencies for this sample of SMEs are listed in Table 2. This metric can be interpreted as the probability that a SME would correctly answer an item. Competencies for this sample range from 0.32-0.76 with a mean of 0.56± 0.11. With a sample size of thirty-six, and average competency level of 0.56, it ought to be possible to correctly classify (as either

“true” or “false”) at least 95% of the items on the "Online Privacy and Security" questionnaire with a 0.999 confidence level [13].

Table 2. Estimates of competency for thirty-six SME’s questioned about "Online Privacy and Security."

SME	Competency	Organization	Location
1	0.48	C2E	M3B
2	0.60	C2E	M3B
3	0.41	C2E	M3B
4	0.56	C2E	M3R
5	0.48	C8E	M3B
6	0.75	ICI	N3X
7	0.75	Missing	Missing
8	0.67	I9B	M2R
9	0.32	C7H	M3B
10	0.58	C8I	N1X
11	0.61	C1B	M3B
12	0.47	C2I	M3B
13	0.52	I0B	M2R
14	0.50	C2I	M3B
15	0.42	C2F	M3R
16	0.52	C8E	M3R
17	0.59	I5I	M2B
18	0.45	C7E	M3B
19	0.59	C8E	M3B
20	0.67	C2I	M3R
21	0.55	C2A	M3B
22	0.67	I5I	M2R
23	0.46	C8F	M3B
24	0.76	C8B	N3Z
25	0.35	I5B	M3R
26	0.51	C8F	M3B
27	0.67	C8F	M3B
28	0.52	I5H	M2R
29	0.72	C7H	M3B
30	0.52	A4B	M2R
31	0.63	Missing	Missing
32	0.58	I9D	M2B
33	0.64	C2A	M3R
34	0.51	I9I	M2B
35	0.69	C2A	M3B
36	0.59	C2A	M3B

4.3 Knowledge Derivation

By using SME competencies as weights, the most probable set of answers can be estimated from SME responses with Bayes' formulation in Equation 4. In Table 3 the answers obtained in this way are compared to the dominant responses given by the 1,482 respondents who completed the GVU survey. Pearson's Chi-square (with Yate's continuity correction) was calculated to test for independence between the two sets of answers [28]. A Chi-square value of 11.852, with one degree of freedom, was obtained from the test, and with a p-value<0.001, there is strong support to conclude that the answers estimated through Consensus Analysis are not different from those obtained for the GVU sample. A Yule's Q= 0.78 also indicates that this association is a reasonably strong one.

Table 3. Cross tabulation comparing majority responses on GVU survey to those estimated from responses of Telcordia SMEs with Consensus Analysis.

	Telcordia SMEs		**Marginal Totals**
GVU Survey	False	True	
False	20	14	**34**
True	5	28	**33**
Marginal Totals	**25**	**42**	**67**

4.4 SME Classification

With estimates of SME competencies in hand, the spatial arrangement of points plotted in Figure 3 can be given a particularly nice intuitive interpretation. Those SMEs who knew the most about "Online Privacy and Security" are plotted in the center of this figure; in fact, those ten SMEs with the highest competency scores fall within the shaded area; while those with the lowest scores are located at the periphery of this plot. However, there also exists idiosyncratic variation among these SMEs in what they know about this domain, and so domain expertise seems to cross organizational boundaries. This idea was tested in several ways.

The terms that SMEs provided to describe their technical areas of expertise were carefully enumerated. Surprisingly, while the frequent use of "hot buzz words" was anticipated, it seems that SMEs exploited free-listing as an opportunity to create very specialized identities. In fact this sample of SMEs applied 189 unique descriptors (each consisting of one or more terms) to characterize their expertise. The number of descriptors listed by SMEs ranged from 0-16 with an average list size of 5.25 descriptors. Four SMEs listed no terms. Only nineteen of the 189 descriptors were listed by more than one SME and all but two of these nineteen were listed only twice, further suggesting a reason for the absence of any discernable clustering of points in Figure 3. However, the five most competent SMEs (24, 6, 7, 29, and 35) identified themselves as knowing more about business and marketing aspects of telecommunications, e.g., "Business planning", "economics", "market-oriented programming"; and used terms such as "system administration" and operations "hand-offs",

implying greater familiarity with consumer or user-oriented perspectives. The only shared concepts expressed in the free lists of those SMEs (9, 25, 3, 15, and 18) with the lowest competency scores were “distributed computing”, “Internet”, “Internet Protocols”, and to some degree more abstract interests, e.g., “mathematics”, “formal methods”, and “theory of distributed systems”. It seems that this group is focused more on privacy and security from a network engineering or design perspective, rather than from a consumer's point-of-view.

More rigorous statistical tests of the organizational and locational basis for knowledge distribution among these SMEs were also made. For these tests, two other symmetrical distance matrices were constructed: the first from SME organization numbers and the second from their office locations (both listed in Table 2). The Organization Code consists of three characters that identify a SME's lab, department and group, respectively. A matrix, whose cells express the organizational distance between SMEs, was constructed from this information in the following manner: a “0” was assigned to all cells along the superdiagonal, a “1” was entered into a cell for SMEs belonging to the same lab, department and group, a “2” for SMEs belonging only to the same lab and department, a “3” to those SME's belonging only to the same lab, and a “4” to those SMEs in different labs. The Office Address also consists of three characters identifying a SME's office site (two possible sites separated by about 50 miles), floor and wing. A locational distance matrix for SMEs was calculated in the following manner: a “0” was assigned to all entries along the superdiagonal, a “1” was entered in a cell for two SMEs located at the same site, on the same floor, and in the same office wing, a “2” for SMEs occupying only the same site and floor, a “3” for those SMEs only located at the same site, and a “4” to those SMEs located at different sites.

The strength of association between these two distance matrices and each of the two distance matrices and the consensus matrix was tested using Quadratic Assignment [29]. With Quadratic Assignment a correlation statistic γ is computed between the corresponding cells in two matrices of observed data. Then one of these matrices is repeatedly permuted randomly, each time computing a new γ. A p-value for this randomization test is determined by counting the proportion of times the value of γ computed for the data permutations equaled or exceeded the value calculated for the observed data. The results obtained from Quadratic Assignment testing with the Consensus matrix, and the Organizational and Locational distance matrices, after 1,000 permutations, are given in Table 4.

Table 4. Results from significance testing of relationships between organizational distance, inter-office distance and amount of consensus among SME responses. (Quadratic Assignment with 1,000 permutations used for tests.)

Association	γ (observed)	Proportion As Large
Organization/Location	0.586	0.000
Organization/Consensus	-0.388	0.810
Location/Consensus	-0.187	0.160

Several conclusions can be drawn from these tests. As one might expect, there does seem to be some association between a SME's organizational affiliation and the location of his/her office. However, there exists little evidence to support the claim that either their

organizational affiliation or the location of their office has much to do with what they know about "Online Privacy and Security," though location does seem to influence more what one knows than organizational affiliation, i.e., a possible "water cooler" effect. Another way of putting this is that cross-organizational forums and informal sharing of information among those who experience greater face-to-face contact may contribute more to learning and knowledge formation than hiring practices and interactions structured more strictly along organizational boundaries.

5. Conclusions

The experimental results obtained for the *Schemer* prototype are promising, especially considering that the "correct" answers obtained for the GVU sample were in many cases tentative due to a large, heterogeneous sample. Moreover, some of these answers were derived statistically, with no prior analysis to weed-out items with near equal frequencies of "true" and "false" responses. This finding implies that meaningful answers to difficult and "fuzzy" problems might be obtained more quickly, and with less effort and cost, from the information provided by a few competent SME's, rather than from a very much larger survey sample [13].

So, what does this experiment have to do with the Semantic Web? We believe that it demonstrates a potentially powerful use of consensus for deriving semantically-relevant ontologies from domain experts. While this experiment asked SMEs to evaluate items pertaining to Internet security and privacy, they might instead have been requested to rate terms in a list on the basis of their salience to a knowledge domain, or to rate the relative strength of semantic relationships between terms on this list. The protocol adopted for the present experiment could be applied to analyze SME responses to these items to determine (1) those terms that should be part of a controlled vocabulary, and (2) a standard set of semantic relationships between terms in this vocabulary. Based on these consensus views other items could be developed and evaluated by SMEs to derive defining attributes for terms in the ontology. At each step in this process, Consensus Analysis provides important "reality" checks. The metrics it yields, as computed in this study, more clearly indicate the saliency of the targeted domain to SMEs and provide an opportunity to assess how much domain knowledge is possessed by each SME in the sample. We conclude with the conjecture that, by interviewing even a small number of competent SMEs, ontologies for Web catalog and directory services can be similarly constructed in a manner that best represents the collective wisdom of semantically-specialized communities-of-interest.

6. Future Work

This study provides motivation for future research in four key areas: Information acquisition, knowledge derivation, knowledge representation and knowledge reuse.

Information acquisition. The ubiquity of the Web is encouraging some in the Knowledge Management community to consider the automation of tried-and-tested information gathering techniques, e.g., repertory grids [30]. Many such techniques exist, and it isn't always clear when application of any particular one is appropriate, e.g., see [31, 32, 33]. Thus, there is a need to consider which of the many available information acquisition

techniques are appropriate for gathering the information needed to derive different types of knowledge, e.g., controlled vocabularies, ontologies or production systems, and how best to deploy them electronically. In fact, a taxonomy of knowledge types, with a mapping of acquisition methods to each, is needed.

Fortunately, this study was able to reuse items developed for the *GVU Survey*. But this was only a proof-of-concept. Any meaningful implementation of the *Schemer* System (or another like it) will require support for item development, preferably by incrementally building pools from items submitted by SMEs. As in test development, items will have to be classified by their author's COI, then carefully pretested and analyzed for their discriminability before being added to an item pool. There is also an opportunity for evaluating alternative protocols for presenting items to subjects based on their background and the knowledge domain being tested, and more flexible highly-interactive formats for presenting items to subjects electronically.

Knowledge derivation. Consensus Analysis provides a rigorous framework for deriving knowledge from information acquired from a group of SMEs. However, further refinements of the method are required to accommodate missing information and guessing, different difficulty levels of problems, and to enable appropriate classification of decision-makers into their respective communities-of-interest. These features are particularly important for supporting the idea of acquiring information from SME's incrementally and at their convenience. With regards to this last point, we envision the use of wireless communication devices, e.g., PDAs, as a useful means to acquire information from SMEs in less-intensive, asynchronous sessions.

Knowledge representation. For derived knowledge to be useful, it needs to be represented in a way supported by other tools, but without sacrificing information about the details of its structure and semantics. Hence, the expressiveness and adequacy of existing knowledge representation standards, e.g., KIF [34], KRSL [35], RDF Core [2] and DAML+OIL [3] need to be reviewed and evaluated.

Knowledge reuse. A minimal use of derived knowledge would be to publish it electronically. However, in the case of some knowledge, e.g., controlled vocabularies or ontologies, new services will be required to support rapid integration of this knowledge with other technology. Thus, there is need to further explore new technologies for making knowledge more accessible to end-users and software applications.

7. Summary

This paper described the *Schemer* prototype, a Web-based infrastructure to acquire information from domain experts and process this information with a quantitative technique known as *Consensus Analysis.* This approach yields metrics that determine (1) whether a particular problem domain has salience for a group of subject matter experts, (2) the level of competency for each of the subject matter experts, (3) the consensus view of this group weighted by the competency of its members, and (4) a classification of subject matter experts by their appropriate community-of-interest.

There is an opportunity to harness the same social-cultural processes that fostered the creation, growth and success of the current Web to evolve rich ontologies. These will be the focal point of the "emergent" Semantic Web and will be constructed dynamically based on

consensus processes. Distributed and semantically-rich information spaces, supported by the infrastructure needed to easily navigate them, promise to be the transforming technology of the 21st century. New knowledge derivation techniques, such as consensus analysis, embedded in tools that enable dynamic evolution of ontologies are a critical component of the semantic Web. We see the *Schemer* prototype as an important step in the long march towards realizing a semantic Web infrastructure.

Acknowledgements

We want to thank Mark Rosenstein, Jon Kettenring, Sid Dalal and anonymous reviewers for commenting on earlier drafts of this paper, and Kim Romney, Sue Weller and Steve Borgatti for many fruitful exchanges on Consensus Analysis and its formal foundations. We are also grateful to Tracy Mullen, Marek Fiuk and Chumki Basu for their assistance with implementing the *Schemer* prototype, and to other colleagues in Telcordia Technologies' *Information and Computer Sciences Research* and *Internet Architectures Research* labs for participating in the experiment described in this paper. Both Insightful *S-Plus*® version 5.0 and Analytic Technologies *AnthroPac*® version 4.1 were used to benchmark data analysis.

References

[1] Fensel, D., 2001. Understanding is based on Consensus. Panel on Semantics on the Web. *10th International WWW Conference, Hong Kong*, 2001.

[2] The RDF Core Working Group. http://www.w3.org/2001/sw/RDFCore/

[3] D. Broekstra et al., 2001. Enabling Knowledge Representation on the Web by extending the RDF Schema. *10th International WWW Conference, Hong Kong* 2001.

[4] T. H. Davenport and L. Prusak. 1998. Working Knowledge: How Organizations Manage What They Know. Boston: Harvard Business School Press.

[5] D. Gibson, J. Kleinberg, and P. Raghavan. 1998. Inferring web communities from link topology. *Proceedings of the 9th ACM Conference on Hypertext and Hypermedia*.

[6] W. Hill, L. Stead, M. Rosenstein, and G. Furnas. 1995. Recommending and evaluating choices in a virtual community of use. *Proceedings of the CHI-95 Conference, Denver, CO.*

[7] C. F. Cargill, 1989. Information Technology Standardization: Theory, Process, and Organizations. Bedford, MA: Digital Press.

[8] C. Cargill, 1994. Prologue and Introduction. *Standard View* **2(3)**: (1994) 129.

[9] A. Farquhar, R. Fikes, and J. Rice. 1996. The Ontolingua Server: A Tool for Collaborative Ontology Construction. Knowledge Systems Laboratory, KSL-96-26 (September).

[10] M. M. Turoff and S. R. Hiltz. 1996. Computer-based Delphi processes. In M. Adler and E. Ziglio (eds.), *Gazing into the Oracle: The Delphi Method and its Application to Social Policy and Public Health.* pp. 56-85. London: Jessica Kingsley Publishers.

[11] R. M. Cooke, 1991. *Experts in Uncertainty: Opinion and Subjective Probability in Science.* New York: Oxford University Press.

[12] H. A. Linstone and M. Turoff (eds.). 1975. *The Delphi Method: Technique and Applications.* Reading, MA: Addison-Wesley.

[13] A. K. Romney, S. C. Weller, and W. H. Batchelder. 1986. Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist* **88(2):313-338**.

[14] R. M. Keesing, 1974. Theories of culture. *Annual Review of Anthropology* **3:73-97**.

[15] A. L. Kroeber, 1948. *Anthropology*. New York: Harcourt, Brace.

[16] D. W. Read and C. A. Behrens. 1989. Modeling folk knowledge as expert systems. *Anthropological Quarterly* **62(3):107-120**.

[17] D. W. Read and C. A. Behrens. 1991. Computer representation of cultural constructs: New research tools for the study of kinship systems. In M. S. Boone and J. J. Wood (eds.), *Computer Applications for Anthropologists*. Pp. 228-250. Belmont, CA: Wadsworth Publishing Co.

[18] FGDC. 1994. *Content Standards for Digital Geospatial Metadata (June 8).* Washington, D. C.: Federal Geographic Data Committee.

[19] B. J. Hillman, R. G. Swensson, S. J. Hessel, D. E. Gerson, and P. G. Herman. 1997. Improving diagnostic accuracy: A comparison of interactive and Delphi consultations. *Investigative Radiology* **12:112-115**.

[20] B. W. Boehm, P. Bose, E. Horowitz and M. J. Lee. 1994. Software requirements as negotiated win conditions. *Proceedings of Information CommunityRE (April)* Pp. 74-83.

[21] R. G. D'Andrade, 1981. The cultural part of cognition. *Cognitive Science* **5:179-195**.

[22] W. H. Batchelder and A. K. Romney. 1986. The statistical analysis of a general Condorcet model for dichotomous choice situations. In G. Grofman and G. Owen (eds.), *Information Pooling and Group Decision Making*. Pp. 103-112. Greenwich, CT: JAI Press.

[23] W. H. Batchelder and A. K. Romney. 1988. Test theory without an answer key. *Psychometrika* **53:71-92**.

[24] D. Caulkins and S. Hyatt. 1999. Using consensus analysis to measure cultural diversity in organizations and social movements. *Field Methods* **11(1): 5-26**.

[25] S. C. Weller and N. C. Mann. 1997. Assessing rater performance without a standard using consensus theory. *Medical Decision Making* **17:71-79**.

[26] A. L. Comrey, 1962. The minimum residual method of factor analysis. *Psychological Rep*orts **11:15-18**.

[27] S. S. Schiffman, M. L. Reynolds, and F. W. Young. 1981. *Introduction to Multidimensional Scaling: Theory, Methods and Applications.* New York: Academic Press.

[28] H. M. Blalock, Jr. 1972. *Social Statistics*. New York: McGraw-Hill.

[29] L. J. Hubert, 1987. *Assignment Methods in Combinatorial Data Analysis. Statistics, textbooks and monographs***, (73)** New York: Marcel Dekker, Inc.

[30] J. H. Boose, 1989. A survey of knowledge acquisition techniques and tools. *Knowledge Acquisition* **1(1): 3-37**.

[31] H. R. Bernard, 1988. *Research Methods in Cultural Anthropology*. Newbury Park, CA: Sage.

[32] J. P. Spradley, 1979. *The Ethnographic Interview*. New York: Holt, Rinehart and Winston.

[33] O. Werner and G. M. Schoepfle. 1987. *Systematic Fieldwork* (2 vols). Newbury Park, CA. Sage.

[34] M. R. Genesereth and R. E. Fikes (eds.). 1992. *Knowledge Interchange Format, Version 3.0 Reference Manual*. Computer Science Department, Stanford University, Technical report Logic-92-1.

[35] J. Allen and N. Lehrer. 1992. *Knowledge Representation Specification Language (KRSL), Version 2.0.1 Reference Manual*. Draft of the DARPA/Rome Laboratory Planning and Scheduling Initiative. ISX Corporation.

Appendix A

Items derived from Georgia Tech's Graphics, Visualization and Usability Center's *10th World Wide Web User Survey* on "Online Privacy and Security." Answers in parentheses based on simple "majority view" obtained from survey of 1,482 respondents. (See *http://www.gvu.gatech.edu/gvu/user_surveys/survey-1998-10/*.)

In general, how concerned are most WWW users about security on the Internet (e.g., others reading their email, finding out what websites they visit, etc.)? Keep in mind that in this context "security" can mean privacy, confidentiality, and/or proof of identity for a WWW user or for someone else.

1. Older (50+ years old) WWW users tend to be more concerned than younger users. (F)

2. Experienced (> 4 years experience) WWW users tend to be less concerned than inexperienced users. (T)

In general, how concerned are most WWW users about security in relation to making purchases or banking over the Internet? Keep in mind that "security" can mean privacy, confidentiality, and/or proof of identity for a WWW user or for someone else.

3. Older (50+ years old) WWW users tend to be more concerned than younger users. (F)

4. Experienced (> 4 years experience) WWW users tend to be less concerned than inexperienced users. (T)

One thing that makes it difficult to study Internet security is people's and business' reluctance to report security problems for fear of causing more problems for themselves. In addition, it is not always clear where they should be reported. One idea is to have a "clearinghouse" where security problems can be studied and tracked. Please provide you opinions about how such an idea might be received.

5. Most WWW users would be willing to report a security break-in of their personal machine or network to a clearinghouse that maintained their anonymity? (T)

6. Most WWW users would be willing to report a security break-in of their business machine or network to a clearinghouse that maintained their anonymity? (T)

7. Less than 10% of WWW users have ever had their credit card number stolen (either online or offline)? (F)

8. More than 50% of WWW users are willing to use their credit card on the web? (T)

9. Less than 20% of WWW users have an unlisted phone number? (F)

10. Most WWW users are unwilling to put their name and address in a directory for public access on the Web (e.g. the online equivalent of a phone company's "White Pages")? (F)

11. Most WWW users are willing to conduct banking on the Web without a statement from the bank of the security procedures used? (F)

12. WWW users within the United States are more concerned than those in Europe or elsewhere about conducting business online outside of their own country without a statement of the security procedures used? (T)

13. In general, PRIVACY is more important than CONVENIENCE to most WWW users? (T)

14. WWW users will more likely participate in an "online auction" for something they are interested in purchasing? (F)

15. Most WWW users think using the Internet for shopping and banking would make their life easier? (T)

16. For most WWW users security features are the deciding factor in choosing whether or not to do business with an Internet-based company? (F)

17. Most WWW users believe that metrics to measure "how secure" a specific site is rated would not be of any help or value to them? (F)

When one views a Web page, they issue a request to a machine that returns the page to them. Which of the following information do most WWW users believe is technically possible to record/log about their page request?

18. Their email address (T)
19. Time of the request (T)
20. Their machine address (T)
21. The requested page (T)
22. An identifier that persists across visits to that site (T)
23. The type of browser they are using (T)
24. Their machine's operating system (T)
25. Their geographical location (F)
26. Their screen size (F)

What information would most WWW users agree ought to be collected for each Web page they request?

27. Their email address (F)
28. Time of the request (T)
29. Their machine address (F)
30. The requested page (T)
31. An identifier that persists across visits to that site (F)
32. The type of browser they are using (F)
33. Their machine's operating system (F)
34. Their geographical location (F)
35. Their screen size (F)

Most WWW users would give demographic information to a Web site ...

36. if a statement was provided regarding what information was being collected (T)
37. if a statement was provided regarding how the information was going to be used (T)
38. in exchange for access to the pages on the Web site (F)
39. in exchange for a small discount at the Web site's store or on their products (F)
40. in exchange for some value-added service (e.g., notification of events, etc.) (F)
41. if the data would only be used in aggregate form (i.e., not on an individual basis) (T)

What conditions would cause most WWW users to refrain from filling out online registration forms at sites?

42. Takes too much time (T)

43. Required to give their name (F)
44. Required to give an email address (F)
45. Required to give their mailing address (F)
46. Information is not provided on how the data is going to be used (T)
47. Accessing the site is not worth revealing the requested information (T)
48. The entity collecting the data is not trusted (T)

Recent attention has been given to mass electronic mailings (a.k.a. spammings) which often contain advertisements, political statements, get-rich-quick schemes, etc. Among most WWW users, which of the following policies would most likely find support?

49. The Government ought to pass a law making it illegal. (F)
50. Mass mailing agencies ought to have to pay an 'impact' fee. (F)
51. A blacklist of spammers should be built to allow message filtering. (F)
52. A registry ought to be created which contains a list of those not wishing to receive mass mailings. (F)

53. Most of the time upon receiving a mass mailing, WWW users will read the message, then either send a message back asking not to be included in future mailings, retaliate in some manner (e.g., mailing bombings, denial of service, etc.), or perform some other action. (F)

Most WWW users would support which of the following?

54. New laws to protect privacy on the Internet. (T)
55. The establishment of key escrow encryption (where a trusted party keeps a key that can read encrypted messages). (T)
56. Web sites need information about their users to market their site to advertisers. (T)
57. Content providers have the right to resell information about its users to other companies. (F)
58. A user ought to have complete control over which sites get what demographic information. (T)
59. Magazines to which a WWW user subscribes have the right to sell their name and address to companies they feel will interest that user. (F)
60. WWW users like receiving mass postal mailings that were specifically targeted to their demographics. (F)
61. WWW users like receiving mass electronic mailings. (F)
62. WWW users ought to be able to take on different aliases/roles at different times on the Internet. (T)
63. WWW users value being able to visit sites on the Internet in an anonymous manner. (T)
64. WWW users ought to be able to communicate over the Internet without people being able to read the content. (T)
65. WWW users would prefer Internet payment systems that are anonymous to those that are user identified. (T)
66. Third party advertising agencies should be able to compile usage behavior across different web sites for direct marketing purposes. (F)
67. There ought to be stricter laws to protect children's privacy than adult's privacy on the Internet. (T)