Information Quality in Information Fusion

Galina L. Rogova
Encompass Consulting
Honeoye Falls, NY
U.S.A.
rogova@rochester.rr.com

Eloi Bosse
Electrical and Computer Engineering
Université Laval
Québec City, QC, G1K 7P4
Canada
eloi.bosse@drdc-rddc.ca

Abstract - Designing fusion systems for decision support in complex dynamic situations calls for an integrated human-machine information environment, in which some processes are best executed automatically while for others the judgment and guidance of human experts and end-users are critical. Thus decision making in such environment requires constant information exchange between human and automated agents that utilize operational data, data obtained from sensors, intelligence reports, and open source information. The quality of decision making strongly depends on the success of being aware of, and compensating for, insufficient information quality at each step of information exchange. Designing the methods of representing and incorporating information quality into this environment is a relatively new and a rather difficult problem. The paper discusses major challenges and suggests some approaches to address this problem.

Keywords: Information fusion, human-system integration, information quality ontology, higher level quality, decision making, quality control

1 Introduction

Information Fusion utilizes a large amount of multimedia and multispectral information coming from geographically distributed sources to produce estimates about objects and gain knowledge of the entire domain of interest. Information to be processed and made sense of includes but is not limited to data obtained from sensor data (infrared imagers, radars, chemical, etc.), surveillance reports, human intelligence reports, operational information, and information obtained from open sources (internet, papers, radio, TV, etc). Successful processing of this information may also demand information sharing and dissemination, and action cooperation of multiple stakeholders such as different national and international authorities, law-enforcement and regulatory agencies, and commercial companies.

Such complex environments call for an integrated human-machine system, in which some processes are best executed automatically while for others the judgment and guidance of human experts and end-users are critical. Automatic processes support human users by affording them inferred object tracks and identities, relations between them, as well as possible current and predicted future states of the environment (situations and threat), with likelihood or plausibility tags assigned to them. In their turn, human users not only utilize the results of automated processes to decisions and actions but also use their experience and flow of observations for providing information to the machine processes (e.g. degree of beliefs, situational hypotheses, utilities, arguments, and preferences). Real-time information exchange in such integrated human-machine system is presented in Figure 1.

The problem of building an integrated fusion driven systems is complicated by the fact that data and information obtained from observations and reports as well as information produced by both human and automatic processes are of variable quality and may be unreliable, of low fidelity, insufficient resolution, contradictory, and/or redundant. The success of decision making in a complex fusion driven human-system environment depends on the success of being aware of, and compensating for, insufficient information quality at each step of information exchange. It is necessary to mention that good quality of input information does not, of course, guarantee sufficient quality of the system output. Quality considerations play an important role at each time when raw data (sensor reading, open source and database search results, and intelligence reports) enter the system, and when information is transferred between automatic processes, between humans, and between automatic processes and humans.

The subject of information and data quality has been receiving significant attention in the recent years in many areas including communication, business processes, personal computing, health care, and databases. At the same time, the problem of information quality in the fusion-based human-machine systems for decision making has attracted less attention. The main body of the literature on information fusion concerns with building an adequate uncertainty model without paying much attention to the problem of representing and incorporating other quality characteristics into fusion processes. There are many research questions related to the information quality problem in the designing fusion-based systems including:

- What is ontology of quality characteristics?

- How to assess information quality of incoming heterogeneous data as well as the results of processes and information produced by users?
- How to combine quality characteristics into a single quality measure?
- How to evaluate the quality of the quality assessment procedures and results?
- How to compensate for various information deficiencies?
- How do quality and its characteristics depend on context?
- How does subjectivity, i.e. user biases, affect information quality?

The reminder of this paper is an effort to establish a conceptual framework in which these questions may be addressed.

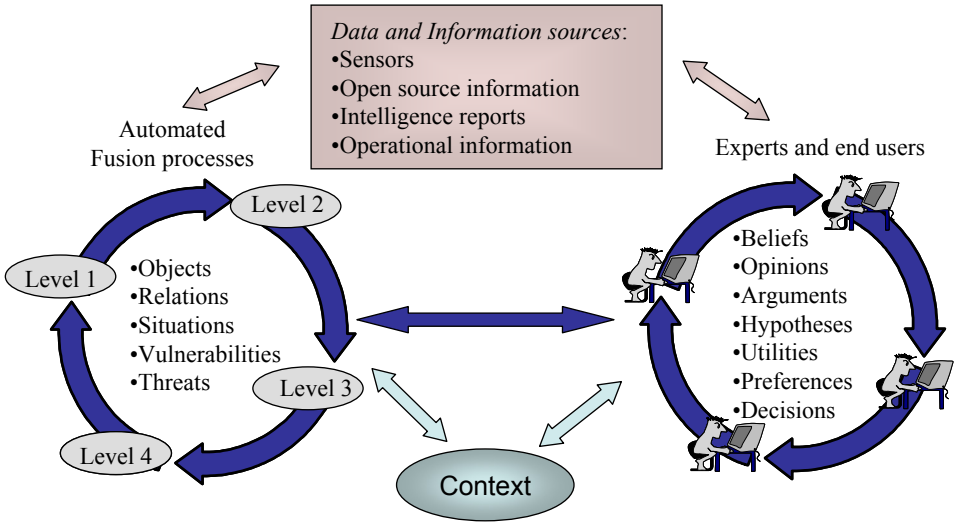

Figure 1 Information exchange in the human-in-the-loop environment

2 Information quality: definition and ontology

There are several definitions of information quality available in the literature:

1. "Quality is the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs." [1]
2. "Quality is the degree to which information has content, form, and time characteristics, which give it value to specific end users." [2]
3. "Quality is the degree to which information is meeting user needs according to external, subjective user perceptions." [3]
4. "Quality is fitness for use." [4]

While having different emphases, all of these definitions point to the fact that information quality is a "user-centric' notion and needs to be measured in terms of the potential or actual value for the users. In the human-system context "users" can be either humans or automated agents and models, and it will be used this way in the remainder of the paper.

Information Quality (IQ) is "information about information," or meta-information, and the best way of representing and measuring the value of this meta-information is through its attributes since "without clearly defined attributes and their relationships, we are not just unable to assess IQ; we may be unaware of the problem." [5].

These attributes have to be considered in relations to specific user objectives, goals, and functions in a specific context. Due to the fact that all users, whether human or automatic processes, have different data and information requirements, the set of attributes considered and the level of quality considered satisfactory vary with the user's perspective, the type of the models, algorithms, and processes comprising the system. Therefore the general ontology designed to identify possible attributes and relations between them for a human-machine integrated system will require instantiation in every particular case.

There have been multiple views on information quality ontologies, identifying quality attributes, classifying them into broad categories and relations. In [3], data quality was classified into four categories: intrinsic,

contextual, representational, and accessibility. In [6], three categories were enumerated: pragmatic, semantic, and syntax while in [5], four sets were identified: integrity, accessibility, interpretability and relevance. In [7], information imperfection, a limited subcategory of IQ, was classified into 2 general categories: uncertainty and imprecision. At the same time there is no clear understanding of what dimensions define information quality from the perspective of information fusion process designers and how different dimensions defining information quality are interrelated. The information quality ontology presented below represents an attempt to fill this gap.

The type of information exchange in the fusion-based human-system environment as presented in Figure 1 notes the three main interrelated categories of information quality proposed in this paper (figure 2):

1. Quality of information source
2. Quality of information content
3. Quality of information presentation

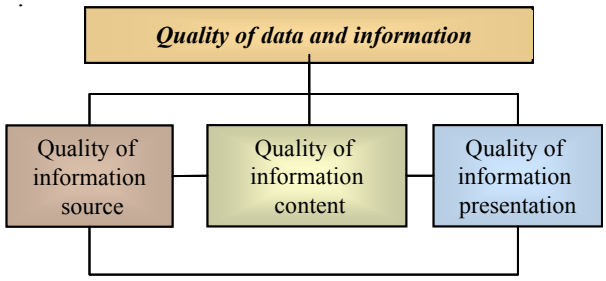

Figure 2. Information quality ontology (a fragment)

The metrics associated with each of these categories are both user- and time-dependent.

The next subsections will consider these quality characteristics in more detail.

2.1 Quality of information content

There are five major attributes of the quality of information content: *accessibility, availability, relevance, timeliness, and integrity.*

Accessibility and *availability* refer to the ability of users to access this information effectively. Accessibility is related to the cost of obtaining this information. Availability is an important characteristic, which has a binary value since information can be either available or not. If availability is 0, all other attributes are irrelevant. At the same time even if information is unavailable now to an agent it doesn't mean it will always be unavailable to all agents. Thus the other attributes still matter to the integrated system.

Timeliness as an attribute of the content of information is different form *timeliness* of information presentation and can be measured by utility of the information under consideration at the time it becomes available.

Data and information can be designated *relevant* if the outcome of the process, decisions, or actions change with the change of the data/information. For example, information can be considered *relevant* for situation assessment if a change of its value affects the set of hypotheses about situations under consideration, the levels of belief assigned to these, hypotheses (e.g. reduces ignorance), or values of utilities of a set of possible courses of action. In the human system environment relevance can be, for example, defined by a human-in-the loop and be represented either by a number between 0 and 1 or in linguistic form (relevant, maybe relevant, irrelevant). *Relevance* of information depends on context as well as goals and functions of decision makers. The dynamics of context, goals, and functions of the decision makers in the dynamic environment make *relevance* a temporal attribute. Thus irrelevant information can become relevant later or relevant information can become obsolete at a certain time [8].

Integrity or lack of imperfection of the content of information is the most studied category of information quality (see, e.g. [5, 7, 9–12]). In the context of a human-system integrated environment imperfection can be defined as something that causes inadequacy or failure of decision making and/or actions. Following [7] we consider two major characteristics of imperfection: *uncertainty* and *imprecision*. Uncertainty "is partial knowledge of the true value of the data" and arises from a lack of information [7]. It can be either objective and represented by *probabilities* reflecting relative frequencies in repeatable experiments or subjective represented by *credibility* (believability) describing information which is not completely trustworthy.

Another uncertainty characteristic, reliability (see, e.g. [9]), can be defined in two different ways. It can be understood as relative stability of information content considered from one time to another under the same environmental characteristics, e.g. sensor readings under the same conditions. It can also be understood as a measure of accuracy of both *probability* and *credibility* and is usually represented by reliability coefficients, which measure adequacy of each belief model to the reality. Incorporation of reliability coefficients is important due to the fact that the majority of fusion operators presume that they are equally reliable. At the same time, it is necessary to account for variable information reliability to avoid decreasing in performance of fusion results.

Imprecision can be possessed by either information with or without error. Thus information without error can be approximate (lacking *accuracy*) or *conflicting* and *inconsistent*. Accuracy represents the degree, to which data corresponds to characteristics of objects or situations. Two latter attributes make sense when either information has several pieces or it is compared with some background information, e.g. databases or information obtained or inferred earlier. *Consistency* of

transient and background information is especially important for situation assessment in the dynamic environment since it can lead to discovery of something new and unexpected critical situations.

Information with error can be *incomplete*, *deficient* (lacking important pieces, which may prevent its usage), vague (ambiguous), or fuzzy (not well defined). The ontology of quality of information content adopted in this paper is presented in Figure 3.

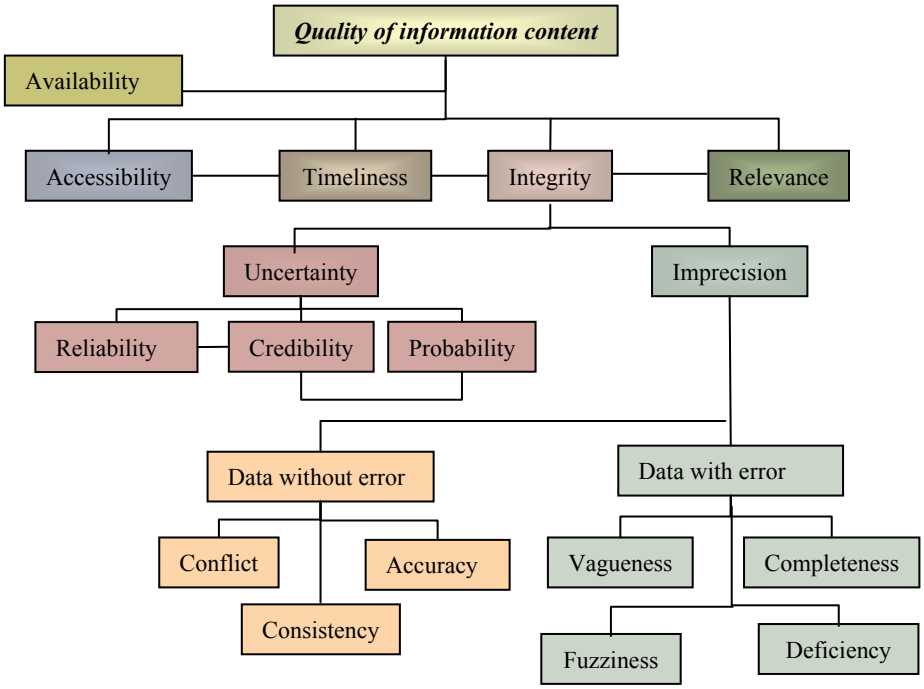

Figure 3. Ontology of quality of information content

2.2 Quality of information sources

From the information quality point of view we consider two types of information sources: subjective and objective. Subjective sources such as human observers, intelligence agents, newspaper reporters, experts and decision makers, supply observations, subjective beliefs, hypotheses, and opinions about what they see or learn. These sources use subjective judgment to produce this information, quality of which is defined by their *level of expertise, reputation*, and *objectivity* as well as their intentions defined by the *truthfulness*. Information coming from subjective sources is usually represented in non-numeric form (so-called "soft information"). Quality of objective information sources such as sensors, models, automated processes is free from biases inherent to human judgment and depends only on how well sensors are calibrated and how adequate models are. As opposite to subjective sources, objective sources deliver information in numerical form (so-called "hard information).

The quality of objective sources comprises *relevance* of the source, which in most cases describes quality of an objective open source, e.g. a source containing objective statistical information as well as their *credibility and reliability*. *Relevance*, *credibility*, and *reliability* can also measure the quality of subjective sources. Example of *credibility* is the frequency, with which a process and model, or a human agent produce a correct answer. *Reliability* is related to quality of beliefs or plausibility assigned to this answer by a human agent or a model [9]. The notion of *reliability* of subjective sources is similar to the notion of *trust* used, for example, in the literature on network centric operations and information sharing (see, e.g. [13]). Ontology of quality of information sources is presented in Figure 4.

2.3 Quality of information presentation

The quality of information presentation affects perception of decision makers and end users, and influences their actions, decisions, judgments and opinions. Information must be presented on time and in a way, which makes it understandable, complete, and easy to interpret. Thus attributes of the quality of presentation are related to when, which, and how information is presented. The ontology of quality of information presentation is shown in Figure 5.

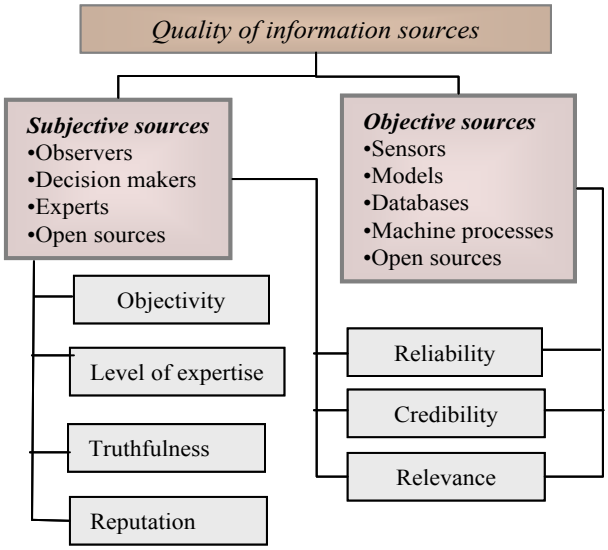

Figure 4. Ontology of quality of information source

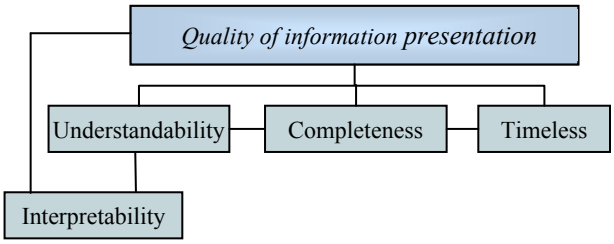

Figure 5. Quality of information representation

Timeliness is affected by two factors: whether the information is presented by the time it must be used and whether the dynamics of information in the real world is reflected by the dynamic of information presentation (when information is presented). *Completeness* is the ability of an information system to represent every meaningful state of the real world system by representing all necessary information related to this state [14]. An important related problem here is whether it is beneficial to present the value of information quality along with information itself [14, 15]. For example, it was found that the benefits of incorporating meta-data depend on the experience of the decision maker [16].

Interpretability defines to what extent the users can understand information presented while *understandability* characterizes the level, to which the user is able to follow the logic of automatic processes producing this information. It is important to mention that the quality of such information attributes as interpretability and understandability depends on the level of training and expertise of the user and can be high for one use and poor for the others.

3 Higher level quality

Higher level quality is "measures how well the quality of information is assessed. The needs for considering the higher level quality stem from the fact that the processes of assessing the value of the attributes of information quality have their limitations. These limitations come from imperfect domain knowledge, difficulty of finding an adequate quality model and its parameters, lack or insufficient amount of additional information that may be required for quality assessment, and subjectivity of quality evaluation performed by experts and end-users. Thus, for example, assessment of probability may need more observations or more time than may be available, and therefore the result is not completely reliable. In some cases point-values probabilities cannot be estimated and are presented by intervals, which represent *accuracy* of probability estimation. Moreover, assessment of overall information quality may require a combination of quality measures expressed in the framework of different belief theories, obtained from decision makers of different levels of expertise, or a combination of quality values represented in both numeric and symbolic form. The information quality ontology shown in figures 2-5 may also serve as the basis for an ontology for higher level quality.

Comprehensive assessment of quality of information requires taking into consideration higher level quality and therefore establishing relations between attributes of quality at different levels. It is important to mention that while quality of source is always a higher level quality for quality of information content, some attributes of information content or information source can serve as higher level quality for other attributes of information source or information content respectively. For example, reliability of a model (a quality attribute characterizing information source) of producing plausibility of a hypothesis is a higher level quality for plausibility assessment (an attribute characterizing information content). We can also consider a certain level of belief that a source of information is reliable as a higher level quality for the assessment of credibility of this source.

4 Assessing the values of information quality

In order to compensate for insufficient information quality in a fusion-based human-machine system it is necessary to be able to assess the values of quality attributes and combine these values into an overall quality measure.

As was mentioned above, the value of quality is user, time and context specific. For example, information about containers arriving in a harbor is not *relevant* when we are evaluating threat presented by a small boat approaching a cruise ship. At the same time, if later we obtain information that the possible attack of a small boat may be a part of a coordinated attack, information

about containers may become relevant. Another example is a *reliability* of an optical sensor, which can be very reliable during a sunny day and not reliable at night. Similar accessibility of a particular piece of information can be considered good if the time required for obtaining this information is much smaller that the time available for making decisions based on this information.

Depending on the context and user requirements the overall quality may relate to a single attribute, a combination of several or all the attributes.

Information quality of single attributes can be assigned by utilizing:

- *a priori* domain knowledge, e.g. sensor signatures, level of training of an observer
- outputs of models and processes producing this information
- supervised or reinforcement learning from examples via, e.g., neural networks
- a level of agreement between sources (e.g., sensors or experts)
- subjective judgment of human users and experts based on their perception.

Selection of a particular method for defining the value of information quality depends on the attribute under consideration and information available. Thus *a priori* domain knowledge about the context of the problem under consideration can provide quality values for many attributes, such as source reputation, information availability and accessibility, or sensor reliability and credibility, which will be difficult or even impossible to estimate otherwise. In many cases *a priori* domain knowledge does not directly contain values of quality attributes but includes certain information such as training examples, which can be exploited for learning. These include, for example, credibility or reliability of sources whether they are human, sensor, or models. Models and processed outputs can serve as a source of assessing integrity of data obtained with these models (reliability, level of conflict, credibility). Subjective judgment of human experts is used when there is no *a priori* domain knowledge of, e. g., probability of non-repeatable events, or when it is important to know the quality of information from the subjective point of view of an expert, e.g. the level of understandability of information.

While the value of any information quality attribute may be defined in various ways, the users and process designers are mostly interested in how good the quality is. Thus the quality scores often measure the level of satisfaction with the information under consideration in relation to the decision makers' goals and objectives or the purpose of the models or processes. Such level of satisfaction with objective quality attributes is based of the value of these attributes and can be expressed either in a linguistic form (e.g., good, fair, bad) or numerically by a number form the interval [0 1]. Some attributes, e.g. availability, have binary values (0 or 1) only. The level of satisfaction then can be represented and treated within a certain uncertainty theory (e.g. fuzzy, belief, or possibility).

One of the ways to measure the level of satisfaction is to measure a particular quality attribute in relation to a certain context specific, and in many cases subjective, threshold (threshold satisfaction). The relation between a threshold and the value of the quality attribute can be transformed into an uncertainty measure within an uncertainty theory under consideration, e.g. beliefs or possibility. In some cases, the quality score of a particular attribute may be defined by comparing a different attribute with a context specific threshold. Thus, for example, if the reliability of source is lower than a certain threshold, information produced by this source may be considered irrelevant. In this case the degree of information relevance can be defined by the function of the distance between the threshold and the source reliability. Subjective quality scores should be considered along with the quality of presentation and the quality of the users and experts.

Evaluation of the overall quality measure requires a combination of different quality attributes. The subset of attributes considered depends on user goals, objectives, and functions as well as the purpose of model or process of interest. In many cases the overall quality represents a trade-off between different attributes, for example, between completeness and understandability, or credibility and timeliness. While designing an overall quality measure one has to take into account the hierarchy of quality attributes, their possible different importance under different context and for different users, and the quality of the values assigned to them (higher level quality). The problem of combing several quality attributes into an overall quality measure is similar to the problem of multi-criteria decision making, which requires comparing several alternatives based on the values of the criteria under consideration. Different quality attributes can be considered and different criteria while alternatives are different values of an overall quality measure.

One of the possible ways of representing such unified quality is to consider a weighted average of the quality scores defined for each component while the weights are non-negative and their sum is normalized to unity. The weights representing a trade-off between the attributes under consideration are context specific and can be assigned by the users based of their needs and preferences. A more general representation of an overall quality measure can be obtained by training a neural network, which can serve as a tool for transforming vector of individual quality scores into a subjective unified quality score. Another way of representing a unified quality of information is to consider utility of decisions or actions or *total information* based on this information. For example, *relevance* can be measured by

the increase of utility of decision or actions after the incorporation this piece of information.

If the quality attribute values are represented within the framework of an uncertainty theory their combination can be obtained by the combination rule defined in this theory. For example, it is possible to use conjunctive combination of reliable quality attributes expressed within the framework of the possibility theory. It is also possible to represent the unified quality measure as a belief network, in which quality of single attributes is expressed within a belief theory [5]. This method of attribute combination is especially appropriate when values of single attributes are heterogeneous, i.e. expressed in different forms e.g. point-value numbers, intervals and linguistic values.

5 Strategies

The success of decision making in a complex fusion driven human-system environment depends on the success of being aware of, and compensating for, insufficient information quality at each step of information exchange. The strategies for quality control can include:

- Eliminating information of insufficient quality from consideration.
- Incorporating information quality into models and processing by:
  — utilizing process refinement by sensor management
  — employing formal methods, e.g. methods of belief change to deal with inconsistency
  — modifying the fusion processes to account for data and information quality.
- Modifying the data and information by compensating for its quality before processing or presenting to the users.
- Delaying transmission of information to the next processing level or to decision makers until it has matured as a result of additional observations and/or computations improving its associated information quality.
- Combination of strategies mentioned above.

Selection of a particular quality control method depends on the type of quality attributes under consideration. For example, *irrelevant* information is usually eliminated form consideration while *credibility* is often explicitly incorporated into models and processes. Methods of compensating for *reliability* are reviewed in e.g. [9]. These methods call for either modifying pieces of information to make them totally or at least equally reliable before processing them or using reliability to modify fusion processes.

Delaying the transferral of information can be used for dealing with such attributes as *reliability, credibility*, *completeness*, or *availability*, quality of which can be improved over time when more information becomes available. Sensor management may be used when it is either impossible or very costly to get information of acceptable quality with current set of sensor or sensor configuration.

Implementation of the quality control measures requires a criterion to be used for defining when the quality of information is not sufficient. One of such criteria is a threshold satisfaction when the value of a quality attribute or a combination of several attributes is compared with a certain threshold. This threshold is highly context specific and depends on decision makers and their attitude toward risk.

The strategy of delaying the transferral of information usually involves incorporation into the overall quality measure such attribute as timeliness. For example dealing with imminent threat requires timely decisions and swift actions. Waiting may result in unacceptable decision latency leading to significant damage and casualties. At the same time, the cost of false alarms can be very high, since it can result in the costly disruption of regular activities, the waste of valuable resources, and scepticism leading to limited compliance when future alarms are sounded. Therefore, the cost of waiting for additional information or the cost of additional computation delay to produce information of better quality and reduce false alarm has to be justified by the benefits of obtaining results of better quality. This can be achieved by either implicitly modeling the expected utility of making a decision based on information of quality at a certain moment or by comparing the quality of information achieved at a certain time with a time varying problem specific threshold [17].

6 Conclusions

The paper discusses major challenges and some possible approaches addressing the problem of data and information quality in the fusion based human-machine information environment. In particular, this paper presents an ontology of quality of information and identifies potential methods of assessing the values of quality attributes, combining these values into an overall quality measure as well as possible approaches to quality control. Designing the methods of representing and incorporating information quality into fusion systems is a relatively new and a rather difficult problem and more research is needed to confront all its challenges.

References

[1] *Standard 8402, 3. I*, International organization of standards, 1986.

[2]J. A. O'Brien, G. Marakas. *Introduction to information systems*, McGraw-Hill/Irwin, 2005.

[3] R. Wang, D. Strong, *Beyond accuracy: what data quality means to data consumers*, J. Management Information Systems, Springer, pp. 5-34, 1996.

[4] J. B. Juran, A.B. Godfrey, Juran's *quality handbook*, 5th edition, McGraw-Hill., New York, 1988.

[5] M. Bovee, R. P. Srivastava, B. Mak, *A conceptual framework and belief-function approach to assessing overall information quality*, International Journal of Intelligent Systems, 18, pp. 51–74, 2003.

[6] M. Helfert, Managing and measuring data quality in data warehousing. In: Proc. of the World Multiconference on Systemics, Cybernetics and Informatics, pp. 55-65, 2001.

[7] Ph. Smets, *Imperfect information: imprecision - uncertainty*. Uncertainty Management in Information Systems: From Needs to Solutions. A. Motro and Ph. Smets editors, pp. 225-254, Kluwer, 1997.

[8]A. Y. Tawfik and E. M. Neufeld, *Irrelevance in uncertain temporal reasoning,* Proc of the 3rd Intl. IEEE Workshop on Temporal Representation and Reasoning, pp. 196-202, 1996.

[9] G. Rogova, V .Nimier, *Reliability in information fusion: literature survey*, in: Proc. of the FUSION'2004-7th Conference on Multisource-Information Fusion, pp. 1158-1165, 2004

[10] P. Bosc, H. Prade, *An Introduction to the fuzzy set and possibility theory-based treatment of flexible queries and uncertain or imprecise databases*, in A. Motro, P. Smets, editors, Uncertainty in Information Systems: From Needs to Solutions, pp. 285-324, Kluwer, 1997.

[11] M. Smithson, *Ignorance and uncertainty: emerging paradigms*, Springer Verlag, 1989.

[12] A-L Jousselme, P Maupin, E. Bosse, *Uncertainty in a situation analysis perspective*, in: Proc. of the FUSION'2003-6th Conference on Multisource-Information Fusion, pp. 1207-1214, 2003,

[13] H. Hexmoor, S. Wilson, S. Bhattaram, A *theoretical inter-organizational trust-based security model*, The Knowledge Engineering Review, Vol. 21 No.2, pp. 127-161, Cambridge University Press, 2006.

[14] Y. Wang, R. Wang, *Anchoring data quality dimensions in ontological foundations*, Communications of the ACM, V.39, No. 11, 1996

[15] I. Chengalur–Smith, D. Ballou, H. Pazer, *The impact of data quality information on decision making: an exploratory analysis*, IEEE Tran. on Knowledge and Data Engineering, V. 11, No. 6, pp. 853-864, 1999,

[16] C. W. Fisher, I. Chengalur–Smith, D. P. Ballou, The impact of experience and time on the use of data quality information in decision making, Information Systems Research, V.14, No. 2, pp. 170-188, 2003.

[17] G. Rogova, P. Scott, C. Lollett, *Distributed Fusion: Learning in multi-agent systems for time critical decision making,* in: Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management, E. Shahbazian, G. Rogova, P. Valen, editors, pp 123-152, FOI Press, 2005.