An Extension of STANAG2022 for Information Scoring

Jérôme Besombes, Adrien Revault d'Allonnes

An Extension of STANAG2022 for Information Scoring

Jérôme Besombes
Office National d'Études et de Recherches Aérospatiales
BP 72 – 29 avenue de la Division Leclerc, 92322 Châtillon Cx
Jerome.Besombes@onera.fr

Adrien Revault d'Allonnes
Laboratoire d'Informatique de Paris VI
104 Avenue du Président Kennedy, 75016 Paris
Adrien.Revault-d'Allonnes@lip6.fr

***Abstract*—We introduce a new confidence scoring method based on an extension of STANAG2022. Our method uses the two parameters included in the STANAG, that is integrates source-trustworthiness to the computation of information-credibility, with two additional parameters: source-proficiency and information-likelihood. These parameters will be formally defined, as will our understanging of the existing criteria. A generic method for calculating a unique score, integrating trustworthiness, proficiency, likelihood and credibility is defined illustrated by two examples: sensor evaluation and information extraction.**

Keywords: Information scoring, sensor evalutation, information extraction.

I. Introduction

Recent evolutions in the tasks of the military and their occupational context (e.g. peacekeeping operations, asymmetric warfare) offer various new challenges in the conception and implementation of information systems. The increasing number of available information makes automatic systems critical. However, the need to manipulate reliable information suggests some caution in the use of completely automated systems. This drawback can be minimised by automatically computing a reliability score for all extracted information. A threshold on this score can then be used to sort information of interest – to be proffered – from the rest.

In order to produce the most comprehensive reliability score possible, it should include different properties in its evaluation: the quality of the source, its ability to produce the information, and the information itself in the light of expert knowledge and other potentialy corroborating or disclaiming information.

II. STANAG2022

The annex to STANAG2022 [1] introduced NATO's intelligence rating scheme. It offered a two-dimensional six-levelled scale (see Table I), wherein confidence in a piece of information was given as a combination of its plausibility and of its source's reliability. The reliability characterises the source, independently of the considered information. Therefore, every information delivered by a source is credited with the same reliability. On the second axis, STANAG2022 attributes the highest level of plausibility to an information that has been confirmed. This suggests that plausibility measures a degree of confirmation: the more an information is confirmed, the higher its plausibility and, conversly, the more an information is condradicted by others, the less plausible it becomes.

This leads us to believe that this rating scheme for information has, at least, the following faults:

1) Using two axes, instead of increasing the readibility of the mark, makes it more obscure. Indeed, which information should be regarded as the most probable between one rated $B3$ and another $C2$? Add to this that the information's plausibility is linked to its confirmation by *distinct* sources and even determining to which of the sources the reliability relates becomes difficult.

2) As noted above, the criteria defined in STANAG2022 to compute plausibility are reduced to the confirmation/denial of the information. Additional criteria could be useful to express the confidence that an information deserves, see [2] for an interesting view on information 'pedigree'. For instance, the context in which two pieces of information are produced by the same source, could be taken into account for the evaluation of the source's reliability, thereby allowing for some flexibility in the source's influence on its production.

These remarks may seem contradictory. On the one hand, we want to give the user a unique score expressing the confidence she should have in some piece of information. In particular, this would simplify comparisons between the confidence in two pieces of information, which is clearly difficult from the two dimensional score of STANAG2022. On the other hand, we'd like to include some criteria not taken into account before, or at least whose randes and definitions appeared meddled in the current system.

To circumvent these difficulties, we will:

- define two additional criteria for scoring an information: source profidency and information likelihood
- integrate each criterion in a process chain which will define a general protocol for calculating a unique score

This work extends the STANAG2022 and provides some real world applications that illustrate how to automate information scoring.

III. Extension of STANAG2022

We will now give our perspective on the existing criteria and introduce the additional criteria we suggest including in our confidence score.

We have chosen to express each criterion on a six-degree scale to conform to the general idea of STANAG2022. Obviously, when taken out of the context of intelligence gathering, or into that of automatic computation, these scales could have more or less degrees, depending on the intended application.

A. Reliability

The reliability of the source is still defined along the original levels (see Table I). However, we limit the range of what this criterion measures [3]. In the available documentation, the AJP-2.1 [4] for one, source reliability not only measures the actual trust vested in the source but also notions of how this source is competent and even some aspects of the confidence it has in its information. Our choice is to split these notions according to their causes and effects, domains and ranges. We define source reliability to be an evaluation of the source independent of all information. This way, this 'opinion' will impact all information emanating from said source in an identical way. Source reliability is not, however, fixed in time. It should and will evolve with time and experience. A reliable source consistently feeding false information will lose some of our trust just as a new source will gradually gain some reliability.

Suppose our source is an electronic sensor. We would argue that source reliability is the capacity of the sensor, in its current state of repair.

B. Proficiency

The capacity of a source to give an information can depend on the information itself. A source may well give an information ouside its realm of expertise. Such a piece of information should not necessarily be disregarded on that basis alone, but the recipient should be aware of this when considering how to classify this information, or how much to believe it.

If we return to our sensor example, for instance, an electronic sensor is calibrated to work in a certain range. This does not prevent it from giving an information outside of this range. In such a context, the scoring should integrate proficiency to modulate the confidence measurement.

Another example is the context of usage. Take meteorological conditions during information collection, for example. They can modify observations for all the data collected while they hold.

We therefore define source proficiency to be a function of the source with respect to the information, rated along the scale given in Table II.

Proficiency	Definition
6	Proficiency cannot be judged
5	Unskilled
4	Insufficiently proficient
3	Partially proficient
2	Proficient
1	Expert

Table II
Proficiency evaluation of a source w.r.t. a specific piece of information

C. Likelihood

Now, suppose we have outside knowledge of the world, stored in an ontology, for instance. By outside knowledge we mean that this knowledge of the world is independent of our information gathering efforts. Suppose that a piece of information is collected which is in contradiction with this general knowledge. We suggest that this contradiction is different from that which may come in the study of our gathered data.

Indeed, if our sensor is, say, a speed detection device. Suppose we train it on a road. Any detected object moving at an impossible speed, knowing the road characteristics, can be considered to be an improbable observation because of our knowledge of the world. This is quite different from comparing two observations at least because our knowledge of the world should not be challenged by the information we are trying to evaluate, whereas any acquired knowledge can be re-rated.

Likelihood qualifies an information based on our global take on the state of the world. Table III shows the degrees we choose to rate it on.

Likelihood	Definition
6	Likelihood cannot be judged
5	Impossible
4	Unlikely
3	Possible
2	Realistic
1	Certain

Table III
Likelihood of the information with respect to outside knowledge of the world

D. Credibility

Because interpretation of the STANAG's initial definition of credibility was influenced by its highest degree, we choose to express it solely as a form of confirmation index. Here again, we must insist on the fact that we are comparing our current piece of information with previously gathered, therefore constructed and rated information, as opposed to

Reliability	Definition
F	Reliability of the source cannot be judged
E	Unreliable
D	Not usually reliable
C	Fairly reliable
B	Usually reliable
A	Completely reliable

Credibility	Definition
6	Truth cannot be judged
5	Improbable
4	Doubtful
3	Possibly true
2	Probably true
1	Confirmed by other sources

Table I
STANAG2022: RELIABILITY OF THE SOURCE AND CREDIBILITY OF THE INFORMATION

outside world truths. Table IV gives the different degrees to which gathered pieces of information can confirm or disclaim each other.

Credibility	Definition
6	Credibility cannot be judged
5	Contradicted by other reliable information
4	Partially contradicted
3	Insufficiently confirmed
2	Partially confirmed
1	Confirmed by other reliable information

Table IV
OUR REVISED CREDIBILITY SCALE: A CONFIRMATION INDEX BETWEEN GATHERED PIECES OF INFORMATION

E. Confidence

The output of our scoring chain is a confidence indicator. It expresses, in a single score, a combination of all the above criteria. One way to read the different levels is given in Table V.

Confidence	Definition
6	Confidence cannot be judged
5	Unlikely
4	Doubtful
3	Possible
2	Likely
1	Extremely probable

Table V
THE OUTPUT OF THE SYSTEM: CONFIDENCE DEGREES IN A PIECE OF INFORMATION

IV. INFORMATION SCORING CHAIN

Now that we have defined the main notions around which our scoring method is articulated, we will outline the essential behaviours of each step in the rating scheme. As we can see, from its illustration in Figure 1, the rating process for a piece of information is a sequential computation. The order in which the rating takes place, as shown in Figure 1, originates in a natural and intuitive decomposition of the process, starting from source evaluation, moving on to a source *and* information criterion and finishing with information appraisal. Figure 1 also indicates the general slopes of impact of each criterion. Each step is a re-evaluation of the previous rating – i.e. the rating from the previous step – when confronted with one of our defined criteria. We will also introduce the idea of different strategies for our criterion evaluation and show how these can be used to model different perspectives on trust and confidence, in a similar way to what we offered in [5]. Concrete examples using these strategies will be detailed in Section V.

Before we delve into the specifics of our method, it should be noted that every evaluation is made on the same scale, i.e. from 1≡'high' to 5≡'low', with 6≡'unratable'. Each evaluation, however, has its own interpretation, given in the previous section. What we are actually building now is the confidence score for a piece of information with respect to the defined criteria.

This confidence score is similar in some ways to existing and commonly used theories. Indeed, what we are offering is an evaluation of uncertain, and possibly imprecise, information. For instance, in [5] we offered a multi-valued perspective on the subject. We therefore think that these theories offer a large amount of tools either for combining our criteria evaluations or even to compute them. As we have said, our six-level scales are a direct consequence of STANAG2022 degrees. Each criterion could, however, be calculated on a continuous range and then be expressed to the user using linguistic variables [6]. The computation of confirmation or conflict could be modelled using Dempster-Shafer theory [7], [8]. Even our confidence level could be integrated to a form of Transferable-Belief Model [9], and associated to its pignistic interpretation.

In all the truth degree tables, the 'unratable' mark is a neutral one. That is, the combination of any confidence score s with a criterion evaluation of 6 will output s. Conversely, the combination of any criterion evaluation e with a current confidence score of 6 will output e or, in the case of source

credibility, e's interpretation with respect to the strategy. This is why Figure 1 shows that any score is reachable from a criterion evaluation of 6. On the other hand, because of its intuitive interpretation of 'the rating cannot be judged', any criterion or piece of information which has evolved away from this mark cannot move back to it. This explains why, in Figure 1, no arrow points from a different rating to 6.

A. Source reliability

Let us consider a new – therefore unrated – incoming piece of information. Because we have never seen it before, we wish to believe it to the extent to which we trust its source. Therefore, the first step in our rating scheme marks an as yet unknown piece of information with the confidence level (Table V, shades of gray in Figure 1) equivalent to the reliability score of its source. Note that Figure 1 shows a 'default strategy' in which A$\Leftrightarrow$1, ..., F$\Leftrightarrow$6. Now, apart for the necessity of the 'unknown' degrees to match – i.e. F$\Leftrightarrow$6 – any other coherent strategy – i.e. where A$\geqslant$B$\geqslant$C$\geqslant$D$\geqslant$E, see Table VI – could be used.

	S_1	S_2	S_3
A	1	2	3
B	2	2	4
C	3	3	4
D	4	5	5
E	5	5	5
F	6	6	6

Table VI
3 DIFFERENT STRATEGIES FOR THE EVALUATION AND USE OF SOURCE-RELIABILITY, STRATEGY S_1 IS THE DEFAULT STRATEGY SHOWN IN FIGURE 1

B. Proficiency and Likelihood

The 'Proficiency' step in our scheme has a negative impact on the score. This is because the initialisation at the source's reliability suggests that the source 'knows what it is talking about', or at least that we trust it in as much as it does. Therefore, the only impact proficiency can have is a negative one. If a trusted source gives us an information it is not qualified to give, we may want to take this with a pinch of salt.

Similarly, the only impact the information's likelihood is likely to have is a decrease in the overall rating. If we believe an information to the amount we believe its source and if the source is qualified to give such an information, the fact that the information is *possible* should not lead us to believe it more, because we initially supposed it to be 'possible', at least. However, if it seems unlikely, we should start doubting it.

Now, to determine by how much either of these two steps will decrease the score, we have to turn to the strategy. Figure 1 shows the general tendency expected for all confidence levels. Once again, it shows the default strategy in which the impact increases in direct proportion to the criterion evaluation. This and some other possible strategies are shown in Table VII.

Since each step of the scoring chain can be seen as a fusion of the previously computed score and a new rating, any probabilistic method [10], [3] or any other fusion operator [11] could also be considered. For instance, conjunctive[1] operators could be used to model a circumspect strategy and, alternatively, disjunctive operators offer a trustful strategy somewhat in the manner of [5]. This choice of operator and policy should be made by the user, depending of the use case.

	S_1	S_2	S_3
1	c	c	c
2	$\min(c+1,5)$	$\min(c+2,5)$	$\min(c+1,5)$
3	$\min(c+2,5)$	$\min(c+3,5)$	$\min(c+3,5)$
4	$\min(c+3,5)$	$\min(c+4,5)$	$\min(c+3,5)$
5	$\min(c+4,5)$	$\min(c+4,5)$	$\min(c+5,5)$
6	c	c	c

Table VII
3 STRATEGIES FOR THE EVALUATION OF SOURCE PROFICIENCY AND INFORMATION LIKELIHOOD, WHERE c IS THE CURRENT CONFIDENCE LEVEL. STRATEGY S_1 IS THE DEFAULT STRATEGY SHOWN IN FIGURE 1

C. Credibility

The last step in our rating scheme is our interpretation of the confirmation index implied by STANAG2022. Because we have divided the notions of information evaluation from that of acquired knowledge comparison, this final step confronts plausible and rated pieces of information. This is an important difference with the previous likelihood step. Indeed, what we consider in the likelihood evaluation step is our general knowledge of the world. We have named this 'outside' world knowledge to distinguish it from constructed – and rated – knowledge. The main point here is that this general knowledge of the world will not be doubted on account of our reasoning. Its confidence level will always be maximal. In this credibility step, however, we will compare pieces of information which we have acquired during our scheme. Each one of these has a confidence level, which may be updated, when compared with new information. Changing one's belief in a piece of information with either corroborations or repudiations is quite natural. However, the distinct notions of confirmations and contradictions themselves imply that this particular criterion has both a positive and negative influence on our confidence level. Because these pieces of information have all been rated with our scheme, our confidence in each of them can be used to weight confirmation (resp. condradiction) computations [12].

D. Feedback

After the credibility step of our scoring chain, the information we were considering has effectively been rated with respect to its source's reliability and proficiency on the information's domain, the likelihood our current knowledge of the world admits for this piece of information and, finally, a confirmation index when compared to currently

[1] A fusion operator is said to be conjunctive (resp. disjunctive) iff the result of the fusion is less (resp. greater) than or equal to the minimum (resp. maximum) of the fused values.

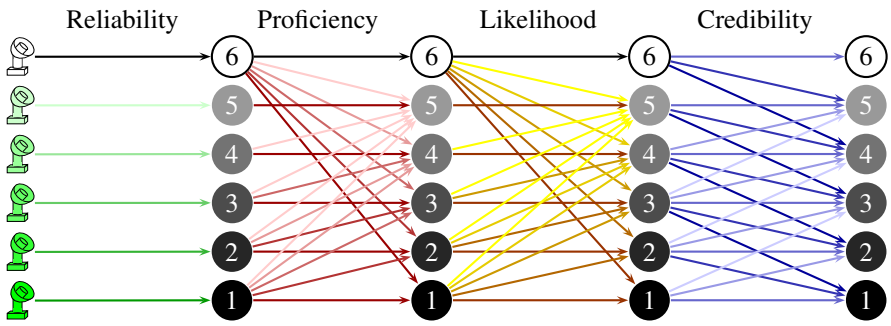

Figure 1. The information scoring chain

uncertain knowledge. At this point, we may consider if we should re-evaluate our trust in the information's source. This feedback of information rating on source reliability is not an integral part of the scoring chain, yet it is central to it. If a source we trust consistently gives us information our system rates as unlikely, we should revise our trust. There are no fixed rules on how this re-evaluation should take place, in part because it depends very much on the source.

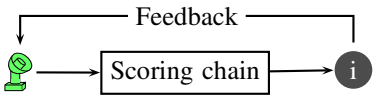

Figure 2. Reliability feedback

V. Applications

We have now introduced our extension of STANAG2022 and all intervening criteria. Table VIII summarises the factors used in the assesment of each criterion. We will now use these notions and processes in two applicative examples to show how and why we mean to use them.

	Reliability	Proficiency	Likelihood	Credibility
Source	✓	✓		
Information		✓	✓	✓
Ontology			✓	
Other information				✓

Table VIII
Evaluation criteria

A. Sensor Evaluation

For our first illustrative example, we will come back to our speed detection device. Suppose we have a brand new, perfectly calibrated tripod-mounted laser device trained on a road. Suppose now that it gives us a reading of a vehicle moving at 80 miles per hour in its detection path. Because we know and trust this sensor in the current conditions of use, we will assign the reading – our piece of information – its high reliability score's equivalent, 1 say.

Now, because the information we got was perfectly in tune with the readings we expect of our sensor – i.e. a speed detection sensor making a speed reading – its proficiency level is at its highest as well. Therefore we will confront our information's current rating of 1 with the source proficiency on its type of 1 also. So after the second step of our scheme, the information saying that a vehicle is traveling at 80 miles per hour on the road which we are watching is still rated 1.

Next comes the information likelihood step. Suppose we have a Geographical Information System providing detailed information on the region in which our sensor is located, and a convincing grasp of the basic laws of motion. Suppose, now, that our sensor is trained on a segment of road right after a tight bend. Our general knowledge tells us that it is highly unlikely that any vehicle can take this bend at a speed in excess of 40 miles per hour and that its speed thereafter is limited to 50 miles per hour by the quality of the terrain. Because this is our outside world knowledge, we place a high amount of confidence on it. We will therefore assume our studied piece of information to be 'unlikely'. When this rating of 4 is fused with the current rating of 1, we will start doubting our information. Depending on the strategy, we may say that our reading is 'possible' or even 'doubtful'.

Finally we have reached the credibility stage. We will need for other information to either corroborate or refute our reading. Suppose we have an agent 10 miles down the road. Suppose also that she spots, within five to eight minutes, a vehicle traveling on the road and we are certain that there is no other vehicle on the road. Obviously, our agent is highly trained and regarded, her information beyond a doubt. Then this new piece of information confirms our doubtful reading and its rating is consequently increased back to an admissible level, 2 say.

The succession of factors in this example is illustrated by Figure 3.

B. Information Extraction Evaluation

In order to deal with the increasing number of textual information (open sources, interceptions and other 'Soft Data'), automatic systems based on named entity recognition can help an operator in his task of extracting information from texts. Such systems are able to extract information from texts automatically, to classify it and to combine it in order to produce more elaborate knowledge. The drawback of such automatic processings is the lack of

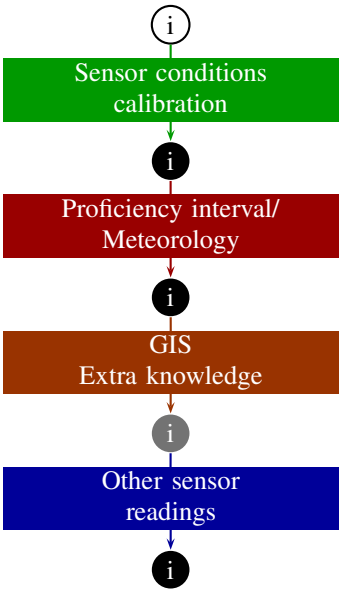

Figure 3. The scoring chain for sensor measurements

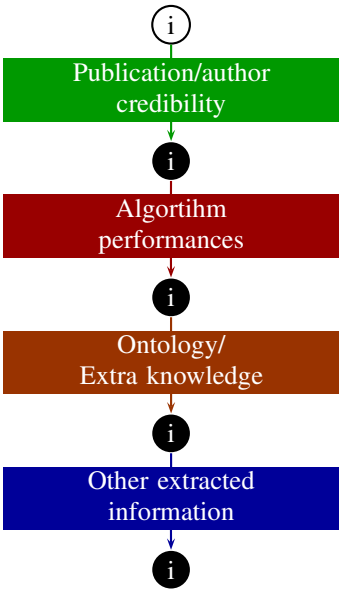

Figure 4. The scoring chain for information extraction

control the user has on the quality of the result. Indeed, the system extracts information and delivers it to the user, for instance as annotations, and the user can then either:

- entirely trust the system and consider that automatically extracted information are always true
- or, remembering that no system is perfect, be suspicious of the automatically extracted information and check it for herself.

Obviously, the second option is impractical and ruled out, since all advantages of using an automatic system would be lost. Blind trust can also be dangerous for any user dealing with critical information.

We will now show how the information scoring chain we defined in Section IV can be applied naturally to add self-evaluation of the produced information to an automatic system.

Suppose we have a textual report produced by a reliable agent and a date automatically extracted from this text (see [13]). The reliability of the agent implies an *a priori* confidence that we can evaluate with a score of 1, say. This initial score marks the first step in the scoring chain.

Given the text and different types of information (a date, a topic, an organisation name, ..., see [14]), a specific information extraction algorithm can be used. Each of these can be associated to an evaluation (recall, precision, f-measure,..., see [15], [16] for an extensive list) that constitutes the proficiency of the algorithm for the information.

Suppose that the algorithm we use has a precision of 90% which is converted to a proficiency level of 2. The choice to favour the precision instead of recall or f-measure, for proficiency evaluation is justified by the fact that the precision measures the probabilty for an extracted information to be relevant, that is, the proficiency of the algorithm relatively to a given information. The extracted date can then be compared with some additional information. Suppose that the event we want to date is a public demonstration in some country and that the initial extracted string was '03/10/08'. The extraction produces the date 'Monday the tenth of March 2008'. If we know that some national holiday in the country of interest, or that the anniversary of an important event falls on the tenth of March, this external knowledge confirms that, perhaps, a demonstration took place on this day. The likelihood of the information is therefore given a score of 1.

Finally, suppose the word 'Monday' is detected further along, in the same text. This information tends to confirm the date of March the tenth, which improves the credibility of the information. Other reports can also contradict or confirm the date.

This instanciation of the scoring chain for information extraction evaluation is illustrated in Figure 4.

VI. Conclusion

Because the context of military intelligence is evolving towards manipulating vast amounts of data, be it structured as with technical sensors or unstructured as with open-source, intercepted or any other form of soft data, we believe that the need for automated systems is increasing. To implement useful systems, one needs to filter out the more informative data to be presented to the user.

Working our way from the original STANAG2022, we offer additional criteria to evaluate a confidence level automatically. We separate the factors used with respect to their ranges and influences and so introduce two new axes in this evaluation and redefine another.

We then explain how we have built an intuitively ordered rating scheme to use our new definitions.

Finally, we illustrate our process, using both a numerical sensor example and a 'soft data' information extraction example, and show how using our definitions offers the user a readable, yet multidimensional estimation of the system's confidence in the considered information.

REFERENCES

[1] North Atlantic Treaty Organization (NATO), "Annex to stanag 2022 (edition 8)," 1997.

[2] C. J. Matheus, D. Tribble, M. M. Kokar, M. G. Ceruti, and S. C. McGirr, "Towards a formal pedigree ontology for level-one sensor fusion," in *Proceedings of the 10th International Command & Control Research and Technology Symposium*, 2005.

[3] G. L. Rogova and V. Nimier, "Reliability in information fusion: literature survey," in *Proceedings of the Seventh International Conference on Information Fusion*, Stockholm, Sweden, 2004.

[4] North Atlantic Treaty Organization (NATO), "Allied Joint Publication – 2.1, Intelligence Procedures," March 2002.

[5] A. Revault d'Allonnes, H. Akdag, and O. Poirel, "Trust-moderated information-likelihood. A multi-valued logics approach," in *Computation and Logic in the Real World, CiE 2007*, Sienna, Italy, 2007.

[6] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning," in *Information Sciences*, 1975, vol. 8, pp. 301–357.

[7] G. Shafer, *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

[8] ——, "Perspectives on the theory and practice of belief functions," *International Journal on Approximate Reasoning*, vol. 4, no. 5–6, pp. 323–362, 1990.

[9] P. Smets, Y.-T. Hsia, A. Saffiotti, R. Kennes, H. Xu, and E. Umkehren, "The transferable belief model," in *Symbolic and Quantitative Approaches to Uncertainty*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 1991, pp. 91–96.

[10] V. Nimier, "Information evaluation: a formalisation of operational recommendations," in *Proceedings of the 7th International Conference on Information Fusion*, 2004, pp. 1166–1171.

[11] I. Bloch, "Information combination operators for data fusion: A comparative review with classification," *Systems, Man and Cybernetics – Part A*, vol. 26, no. 1, pp. 52–67, January 1996.

[12] J. Besombes and L. Cholvy, "Information fusion: using an ontology to information evaluation," in *Proceedings of the International Colloquium on Information Fusion 2007, Xi'an, China*, 2007, pp. 416–422.

[13] J. Makkonen and H. Ahonen-Myka, "Utilizing temporal information in topic detection and tracking," in *Proceedings of 7th European Conference on Digital Libraries (ECDL 2003)*, T. Koch and I. T. Solvberg, Eds. Springer-Verlag, 2003, pp. 393–404.

[14] *Sixteenth Text REtrieval Conference Proceedings (TREC 2007)*. Gaithersburg, Maryland: National Institute of Standards and Technology, November 2007.

[15] *Message Understanding Conference 7 (MUC7)*. Philadelphia: Linguistic Data Consortium, 2001.

[16] National Institute of Standards and Technology, "Automatic Content Extraction Evaluation (ACE07)," May 2007.