Fusion of Open Source Information

David F. Noble
Evidence Based Research, Inc.
1595 Spring Hill Road, Suite 250
Vienna, VA 22182
USA
noble@ebrinc.com

Abstract: *Open source information on the Internet can contribute significantly to such assessments as competitive intelligence, business trends, or evolving social attitudes. Unfortunately, this information is mostly unstructured text, and varies widely in accuracy, focus, and level of impartiality. Estimates on factors important to planning and decision making, such as likely responses of competitors to a corporation's market actions, can be inferred from competitors' past actions, their corporate strengths and interests, and their executives' personalities. Elements related to these factors can be gleaned from press releases, historical records, and corporate profile information. To make the overall assessment, these elements must be combined. This paper describes a methodology for fusion of open source information now under development for a major U.S. government organization. This process employs formal ontologies for the fusion domain and for evidential reasoning, a commercial tool for text extraction and structuring, and tools to help operators review, edit, and augment the fusion products. It features a fusion pedigree to document the audit trail of sources and processes contributing to a fusion product. The fusion steps are: 1) collect structured and unstructured information related to the entity or event of interest; 2) extract and structure free text; 3) create "event reports" from each structured record, whether derived from free text or previously structured sources; 4) create ontology-based communication reports from these records, 5) associate these reports, 6) cue manual search for additional information, and 7) fuse the information.*

Keywords: open source, unstructured information, fusion.

1 The Open Source Fusion Problem

The introduction to a paper on assessing the reliability of open source information [1] summarized the importance of open source information as follows: "In recent years open source information has become increasingly valued. The huge proliferation of open source information on the Internet, including news sites, discussion boards, and chat rooms, often provides the initial reporting and early indicators of important events and activities. Because organizations use the Internet to advertise their capabilities and alliances, these sources help analysts to understand the current competitive landscape and to forecast possible alternative futures. In addition, by providing a common sharable context among analysts, open source information reduces intelligence 'stovepiping.' Therefore open source intelligence often serves as the foundation of information utilized in planning and targeting other high value collection activities. In national intelligence, it provides an important supplement to HUMINT, SIGINT, MASINT, and other more classified collection means. By combining these sources, analysts can understand the diversity of viewpoints on important issues."

That paper then summarized the difficulties of fully taking advantage of open source information: the difficulty of finding key information because the Internet is vast and only a small fraction is indexed; the difficulty of extracting and combining key information because most of the reports are free text; and the difficulty of assessing report credibility and trustworthiness because anyone can post anything on the Internet and because it is difficult to trace the pedigree of posted information. That paper focused on assessing report credibility. This one addresses the second of the obstacles: the difficulty of extracting and combining the information.

Like text extraction, fusion of open source information converts unstructured information into well-defined structured records. Fusion extends text extraction by combining the structured records from multiple sources to create a more complete and precise fused product. In addition, the fusion can also combine the text extractions with structured records from databases or with hand-entered data from analysts.

The approach described here builds on advanced commercial technology for text extraction. Unfortunately, this is still currently a somewhat shaky foundation [2] since commercial information extraction tools still require considerable work to set up, and because it is not always possible to achieve the desired level of recall and precision. As will be described, despite these limitations, with some manual augmentation open source fusion can be accomplished.

2 Example of Open Source Fusion

The following example illustrates the conversion of the information in multiple unstructured sources into a single fused structured record. The example is based on the work EBR is now performing for a client. We have modified the material somewhat, both to simplify the example and to make it more instructive.

In this case, the client is interested in understanding commercial relationships among companies within the telecommunications industry. Among their interests are various technical and marketing associations. In this example, there are two sources, PWID 1 (Published Works #1) and PWID 2. Each source contains multiple references to the nature of the association, with each reference contributing some information.

PWID 1

Indonesia's Telkom awards Ericsson with broadband contract

Xinhua

11/11/2004 15:01

Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services. Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.

"There is already a very small number of lines—lower than 1, 000—available in Surabaya," PT Ericsson Indonesia President Mitch Lewis was quoted as saying.

He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.

Telkom launched its broadband service, called Speedy, in July and is aiming to provide 40,000 connections in Jakarta, and 10,000 in Surabaya in the first phase of the project.

It has allocated US$15 million for the service's infrastructure, or between US$290 and US$300 per line.

PWID 2

Ericsson wins deal in Indonesia

11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers.

The broadband service will include DSL access that will deliver data, voice, and video simultaneously in East Java. Ericsson's work for this contract is slated to begin in April of 2005.

PT Telkom is Indonesia's largest telecommunications company and currently offers a broadband service, called Speedy, which it launched in July of 2004. State-run Telkom plans to convert its network's infrastructure broadband capacity from fewer than 1,000 lines to 2 million lines by 2008.

Fig. 1. Text Input to Fusion Process

The fusion process extracts and combines the information in these two sources to create the following single record (Figure 2):

Name	
OrgAssocType	Technology Partnership
OrganizationsAndRoles (1)	
Organization	Ericsson
Role:	Vendor
OrganizationsAndRoles (2)	
Organization	PT Telkom
Role:	Buyer
AssocStartDate	
Min estimate	April 1, 2004
Max estimate	April 15, 2004
Confidence	High
LocationOfWork	
City	Surabaya
Region	East Java
Nation	Indonesia
Technology	Broadband
Contract Amount	
Estimate	$7.5 million
Lower bound	$7.5 million
Upper bound	$7.5 million
Confidence	High
Communication Reports	Report 1; Report 2; Report 3, Report 4

Fig. 2. Product of Open Source Fusion

All of the information in the structured record can be found dispersed within the sources. The goal of open source fusion is to find it, structure it, and combine it.

Interestingly, the fusion of open source information seems to be Level 1 fusion as defined by the Joint Directors of Laboratories Fusion Panel. As will be described, this fusion process entails the usual steps of level 1 fusion: obtain sensor report, align report data, condition sensor report using sensor model, estimate association probabilities among reports, associate applying strategy for managing uncertain associations, and fuse and update state [3], [4], [5].

In addition, open source fusion requires a "collection management" process in which more coarse-grained broad coverage sensors cue more fine-grained narrow coverage ones. In our case, the course grained sensor is Aerotext™, a state of the art text extractor from Lockheed Martin. Aerotext™ can search vast amounts of information quickly, extracting and structuring with a precision that greatly facilitates the information fusion process. The fine-grained sensors are people, who cannot search quickly but can extract many more details from the text than the automated text extractor can.

3 Fusion Environment and Tools

The fusion process to be described requires an information processing infrastructure for collecting, structuring, and managing information. Figure 3 depicts the major components of the system that EBR uses to collect and structure open source information [6]. At EBR we use primarily Kapow™ to collect open source information, converting Web pages into plain text. We then use Lockheed Martin's AeroText™ to structure the information in the text, creating fully structured records with all fields defined in terms of a formal ontology.

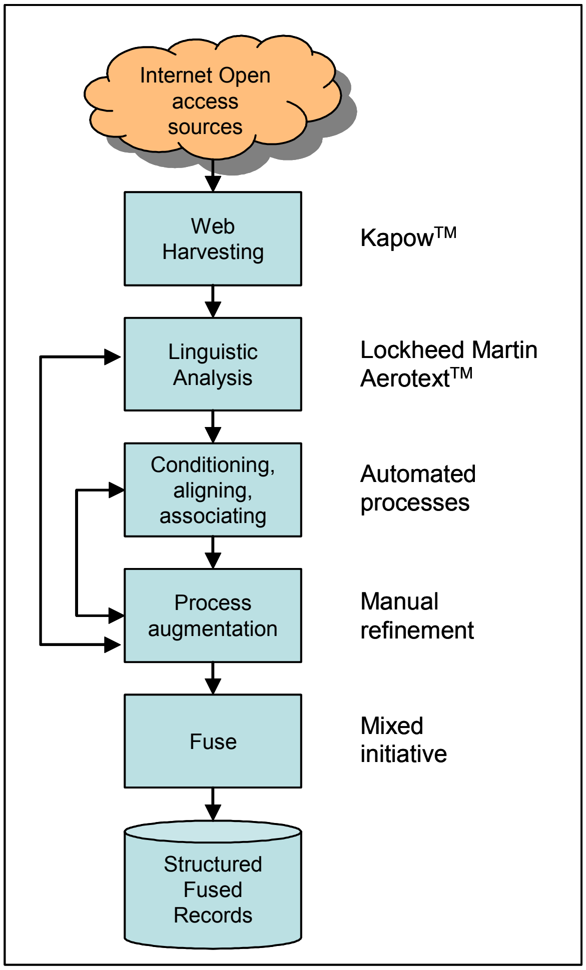

Fig. 3. Open source collection, structuring, and fusion tools

The immediate product of the text extraction and structuring are Aerotext™ "events." These events are analogous to raw sensor reports. These raw reports need to be converted into more processed reports, analogous to the track reports in tracking. To create these more processed reports, the Aerotext™ output needs to be aligned and conditioned. Here, the alignment is enforcement of a common vocabulary defined in a formal ontology. As in sensor fusion, the conditioning needs a sensor model, which in this case is a model of the text extraction tool. This conditioning permits uncertainties to be associated with some of the data fields. These uncertainties then assist in association decisions.

While the goal of the fusion methodology is full automation, at the present time the process requires considerable manual intervention. This manual refinement improves extraction, conditioning, alignment, combination, and state estimation.

The remainder of this paper will illustrate each of these steps for the problem presented in section 2.

4 Open Source Fusion Steps

This paper describes seven steps in open source fusion: 1) process set up, 2) information collection and extraction; 3) conversion of Aerotext™ events into ontology-defined aligned and conditioned "communication" records, 4) manual refinement of the communication records, 5) association of these records; 6) cued manual search for further refining information in the source text, and 7) fusion and state update.

4.1 Process set-up

Process set-up is the work required to prepare the system for fusion. It includes defining the ontology for the domain being examined, setting up Aerotext™ to extract the information defined in the ontology, and generating alignment rules.

In its open source collection and structuring work, EBR employs an ontology based on the Suggested Upper Merged Ontology (SUMO), an effort within the IEEE SUO working group to create a high level ontology for use by expert systems within a variety of domains[6]. We use the OWL file structure and the Protege application to create and edit the ontology.

Our adaptation of SUMO has four principal components: people and organizations, competitive intelligence, telecommunications, and evidential reasoning. The record shown in Figure 2 is a simplification of the "Organization Association" class in the ontology. The ontology-defined analogue of track reports are the "Communication" classes. The subclass employed in this article is called "Organization Association-Communication." These communication classes are the buffer between the immediate Aerotext product (which are not constrained by the ontology) and the actual product classes (like "OrganizationAssociation") which analysts examine to understand the collected information.

Whenever we are tasked to collect information to examine a new domain, we examine the ontology to determine if it includes all of the needed concepts, If not, then we extend the ontology.

The second step in the process set-up is creating the Aertotext™ extraction rules. These rules describe for Aerotext™ how to identify and structure the information to be extracted. In effect, they create fairly abstract templates that describe all the different ways a concept can be expressed in the target language. In our example, these are the different ways that one can express in English that two organizations are entering into an association. Adapting rules to accommodate new issues in a previously examined domain can often be accomplished quickly. Developing rules for a new domain can be labor intensive, sometimes requiring more than a month of effort from experienced Aerotext™ users.

Our Aerotext™ rules specify what to look for within an English sentence and how to represent the information found as structured event records. That is, the rules tell

Aerotext™ how to recognize the issues of interest to be found in free text, and how to deposit the information contained in a sentence into a structured record.

The third issue in set-up is describing to the system how it is to align the Aerotext™ output records into the ontology-defined fusion database. This is required because the vocabulary demands of Aerotext™ for text extraction and of the ontology for well-defined concepts conflict. Aertotext™ wants to know all the different words that can be used to express an idea because that helps it detect and structure these ideas. The ontology, on the other hand, wants to use only a single word to describe a concept, because it wants to ensure that, to the extent possible, when words in the database are different, they define different entities or events, and when they are they same, they define the same entities and events. This characteristic of the ontology is very helpful in the association step of fusion for it helps the system judge whether or not two extracted text records are referring to the same or different things.

4.2 Information collection and extraction

Once the set-up is completed, the system is ready to begin its fusion of open source information. The first step is to collect material that may contain useful information, and then to identify, extract, and structure the information. As indicated in Figure 3, EBR's principal tool for automated collection is Kapow™. Kapow™ uses intelligent spider technologies to auto-harvest / extract information of interest off the web and then saves this information in a database. The next step is for Aerotext™ to find and extract the relevant information. In this case, there are five sentences that contain the information that contribute to the final fused product shown in Figure 2. As noted in Figure 4, Aerotext™ finds four of them, and converts these four to the structured Aerotext™ event records.

PWID 1

Event 1. Region = 24 – 38

Region Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services.

Event 2. Region = 40 – 55

Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.

Missed event

He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April.

PWID 2

Event 3. Region 15 – 37.

11/11/2004—Today, Sweden-based Ericsson was awarded a contract with an estimated value of $7 million to provide broadband technology and related services to PT Telkom's customers.

Event 4. Region 62 – 75.

Ericsson's work for this contract is slated to begin in April of 2005.

Fig. 4. Information Extracts

Aerotext™ labels each of the extracted sentences with an identifier, consisting of the published works ID and the region within the published work. The region is a binary (a,b,) where "a" is the number of "tokens." (basically, words) of the first word from the beginning of the source and the "b" is the number for the last word in the extraction. As will be shown, this location labeling is very important for associating events with each other.

Aerotext™ converts each of these extractions into a structured event record, as shown in Figure 5.

	Event 1	Event 2
Source	PWID 1	PWID 1
Region	24, 38	40, 55
Event type	Contract Event	Technology Event
Text	Telkom has awarded Ericsson with a contract to.....	Ericsson will be providing customers in Surabaya, East Java...
Subtype	Technology	
Organization 1	Telkom	Ericsson
Organization 2	Ericsson	
Contract_Place		Surabaya, East Java
Contract_Amount		
Contract_Date		
Technology	broadband	Ethernet, DSL

	Event 3	Event 4
Source	PWID 2	PWID 2
Region	15, 37	62, 75
Event type	Contract Event	Contract Event
Text	Ericsson was awarded a contract with an estimated value of $7 million to ...	Ericsson's work for this contract is slated to begin in April of 2005
Subtype	Technology	General
Organization 1	Ericsson	Ericsson
Organization 2	PT Telkom	
Contract_Place		
Contract_Amount	$7 million	
Contract_Date		in April of 2005
Technology	broadband	

Fig. 5. Structured Aerotext™ Records

Note that in our use of Aerotext™, the unit of extraction is one sentence. Therefore, information that spans multiple sentences will be represented in multiple structured events.

In this example, there are six desired pieces of information about the association: 1) the organizations involved; 2) the roles of each of the organizations; 3) the date the association begins; 4) the location of the work; 5) the size of the award; and 6) the technology involved (broadband). Each of the extracted sentences has some of the information. The first and third records state both organizations and roles, the association date is in event 4 (and the missed event), the location is in event 2, the size of the award is in event 3 (an estimate) and in the missing event (confirmed amount), and the technology is in events 1, 2, and 3. In addition, the missing event, after it is found, can refine the estimates of contract size in event 3 and start date in event 4. The purpose of the fusion is to create a single record (Figure 2) that combines all of these data and also refines the award amount and start date.

Aerotext™ in finding and structuring these sentences provides an enormously important function, since it can quickly process very large amounts of material that would be uneconomical for people to do. However, comparing the initial material with the text shows some of the limitations of our extraction (better rules would extract better, but they take time to create). First, Aerotext™ missed one of the key sentences. It lacked the specific information that our rules required for finding an event. Second, it was unable to distinguish between the organization roles of PT Telkcom and Ericcson, since the rules didn't specify how to distinguish between vendor and buyer. Third, it used two different labels for PT Telkom: "PT Telkom" and "Telkom." Fourth, the third event did not note that the value of the contract was "estimated." Fifth, the contract start date in event 4, "in April of 2005" remained a string rather than being interpreted as a date. In the next two steps, which converts the Aerotext™ events into ontology-defined "communications" records, all except the first of these issues is addressed.

4.3 Conversion to "communication" records

This step is where the conditioning and alignment of the Aerotext™ events are started. This is the step that sets the stage for associating the five records (after the fifth is found) and then creating a single fused record that contains all of the data. Also, as noted previously, this is analogous to creating track reports from sensor data.

Figure 6 shows the "communication" records created from the first two event records of Figure 5. The communication record is defined as part of the ontology (unlike the Aerotext™ events), and closely resembles the fusion product "organization-association" record of Figure 2. The system automatically created these two records. Other than being in the structure suitable for fusion and creation of the organization record, not much processing has occurred. The system was able to do some alignment, changing "Telkom" in the Aerotext™ event to "PT Telkom," and parsing the location of the work. Both of these alignments will help with the association. The system was not, however, able to do automated conditioning. It did not, for example, attempt to define the level of uncertainty in the date estimate (in report 4) or size of the contract.

Moreover, this automatic processing did not address most of the problems noted in Section 4.2 This will require operator intervention.

Name	Report 1	Report 2
OrgAssocType		
Organizations(1)		
Organization	PT Telkom	Ericsson
Role:		
Organizations (2)		
Organization	Ericsson	
role		
AssocStartDate		
Lower est.		
Upper est.		
LocationOfWork: City Providence Country		 Surabaya East Java Indonesia
Technology	Broadband	Ethernet, DSL
Contract Amount		
Estimate		
Lower bound		
Upper bound		
Confidence		
Pedigree		
Text	Telkom has ...	Ericsson will be
Processor	Aerotext	Aerotext
Source ID	PWID 1	PWID 1
Region	24, 38	40, 55

Name	Report 3	Report 4
OrgAssocType		
Organizations(1)		
Organization	Ericsson	Ericsson
Role:		
Organizations (2)		
Organization	PT Telkom	
role		
AssocStartDate		
Lower est.		04/01/2005
Upper est.		
LocationOfWork: City Providence Country		
Technology	broadband	
Contract Amount	$7 million	
Estimate		
Lower bound		
Upper bound		
Confidence		
Pedigree		
Text	Ericsson was	Ericsson's
Processor	Aerotext	Aerotext
Source ID	PWID 2	PWID 2
Region	15, 37	62, 75

Fig. 6. Ontology-based Communication Records

4.4 Manual refinement of the communication records

The rules that EBR developed for this extraction project are not detailed enough to extract the additional information in the text. Given the current state of the art in automated text extraction, the EBR Aerotext™ team did not feel that it was economical to generate the more complicated rules that might be able to extract this information. Accordingly, EBR instead is creating an environment to facilitate manual entry of the additional data. Figure 7 summarizes the communication records after operator editing. Operator changes are in italics. Note that the operators inserted the organization roles, easily found in the first and third events. Operators also represented the two uncertainties. In order to maintain an audit trail describing the evolution of the data, the operators who updated the records inserted their initials into the pedigree information.

Name	Report 1	Report 2
OrgAssocType		
Organizations(1)		
Organization	Ericsson	Ericsson
Role:	*Vender*	
Organizations (2)		
Organization	PT Telkom	
Role	*Buyer*	
AssocStartDate		
Lower est.		
Upper est.		
LocationOfWork: City Providence Country		 Surabaya East Java Indonesia
Technology	Broadband	Ethernet, DSL
Contract Amount		
Estimate		
Lower bound		
Upper bound		
Confidence		
Pedigree		
Text	Telkom has ...	Ericsson will be
Processor	Aerotext, AJJ	Aerotext AJJ
Source ID	PWID 1	PWID 1
Region	24, 38	40, 55

Name	Report 3	Report 4
OrgAssocType		
Organizations(1)		
Organization	Ericsson	Ericsson
Role:	*Vender*	
Organizations (2)		
Organization	PT Telkom	
role	*Buyer*	
AssocStartDate		
Lower est.		*04/01/2005*
Upper est.		*04/30/2005*
LocationOfWork: City Providence Country		
Technology	broadband	
Contract Amount		
Estimate	*$7 million*	
Lower bound	*$5 million*	
Upper bound	*$9 million*	
Confidence	*moderate*	
Pedigree		
Text	Ericsson was	Ericsson's work
Processor	Aerotext, AJJ	Aerotext, AJJ
Source ID	PWID 2	PWID 2
Region	15, 37	62, 75

Fig. 7. Ontology-based Communication Records

4.5 Associate reports

Deciding whether reports are actually reporting on the same incident is the report association problem, which for many fusion problems is the principal challenge. Reports that refer to the same entity or event can be combined to estimate more accurately the characteristics of the entity or events. Reports that reference separate events or entities cannot be directly fused. When two reports may or may not be about the same event or entity, then the fusion logic needs to decide what to do. It can tentatively combine them, with the provision that the reports can be disassociated if necessary, can set them aside with the provision that they can be combined later, or can generate "multiple hypotheses" in which the reports are combined in one hypothesis and not the other. The latter can lead to very complicated hypothesis management logic.

In fusion of open source information, association decisions are based on two criteria: 1) the consistency and similarity of the content of the data in the reports and 2) the proximity of the text "regions" of the source material.

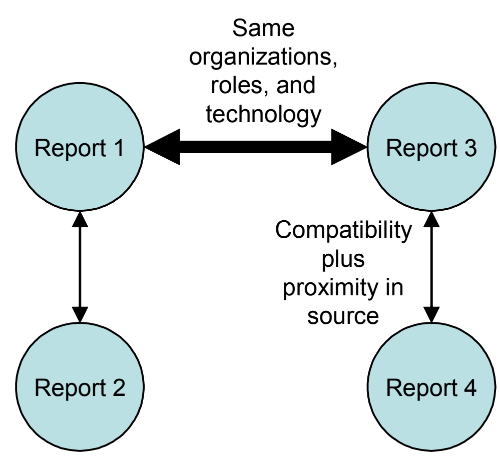

Fig. 8. Report Associations

As summarized in Figure 8, in this example, Reports 1 and 3 are associated based on content similarity and consistency. They are both about "broadband" association with the same organizations in the same role. This amount of match is sufficient to associate the reports, even though they are drawn from different sources.

In contrast, reports 1 and 2 cannot be matched based on the similarity of content. They are consistent, since both concern Ericcson and broadband technology ("broadband" and "Ethernet DSL" are a match since our ontology knows that Ethernet DSL is a kind of broadband, but an operator can also make this judgment). This degree of match by itself would not be sufficient to associate the reports. However, these two reports are adjacent sentences in the original text, as is determined from the source ID and region fields in the pedigree part of the reports. Because adjacent sentences frequently are discussing the same subject, the fact that report 2 directly followed report 1 in the text and the compatibility of the content is sufficient to make the match. Similarly, reports 3 and 4 are matched.

EBR has not yet formalized general rules for report association, and is searching for a proximity measure. This proximity would need to reflect the probability that two reports reference different events and entities given the degree of similarity in their fields. Establishing general criteria for report association, however, a very significant one, for incorrect associations can completely invalidate conclusions based on the association. Such events occasionally get into the news, as was the case when an infant with the same name as someone on a terrorist watch list was not permitted to board a plane.

In the meantime, we do not plan to associate reports that could plausibly be about separate events or entities. This prevents the kind of embarrassing mistake cited above. In addition, deferring association does not undermine the subsequent analysis very much since we use additional techniques to examine our data post-fusion.

4.6 Cue manual search

In examining the reports on the Ericsson–PT Telkom association, an operator might note that the size of the contract is not confirmed, but is only an estimate. Knowing that the automatic text extraction can miss important material, an operator might decide to manually inspect the source material that the other reports are drawn from. As noted previously, this is analogous to a broad view sensor cuing a fine-grained one in surveillance, with in this case Aerotext™ serving as the broad view sensor and people serving as the fine grained one. Once a person reads the source material in "published works 1," he or she will note the missed sentence:

"He confirmed the value of the contract to be $7.5M, with a target start date of late March or early April."

The operator can incorporate this material into the process by manually creating an association communication record. He can specify that the contract amount is $7.5M, and that this number as very high confidence. He can also note that the start date has an early estimate of March 15 and a late estimate of April 15.

4.7 Fusion and state update

The final step in this example is to combine the information in the five records (including the manually produced one) into a single record. In this example, there are three combination rules: 1) if all reports agree on a value, then put that value in the fused product; 2) if some reports cite a value and others do not contradict, then insert those values; 3) if reports reference compatible values, insert the one the ontology uses; 4) if reports cite different numerical value with uncertainty estimates, then update using normal state estimation methods.

The result of applying these fusion rules is the fused product shown in Figure 2.

As part of its "evidential reasoning" ontology, EBR has included several different ways for representing uncertainty. One of the most useful for open source fusion is a probability list of nominal data. Thus, for example, if two reports give different estimates of an entity identity, the fused product is just a list of these two estimated, each associated with a probability.

5 Conclusion

Because of the growing volume and diversity of readily obtained information on the Internet, open source information is becoming increasingly important. In theory, it should be possible to use open source to create an information landscape that summarizes the range and credibility of viewpoints about a vast number of issues. In practice, this is hard to do because of the difficulty of finding key information, extracting and structuring free text, and combining non-numeric information. Today, new tools make it feasible to efficiently collect and structure the needed information. As described in this paper, these same tools now also make it possible to combine information from multiple sources to create a much more complete and precise information product.

6 References

[1] David Noble "Assessing the Reliability of Open Source Information" Proceedings of the 7th International Conference on Information Fusion. Stockholm, Sweden, 2004.

[2] Sarah Taylor, "Improving Analysis with Information Extraction Technology." Proceedings of the 8th International Command and Control Research Symposium. National Defense University, Washington, D.C., 2003.

[3] Edward Waltz and James Llinas. Multisensor Data Fusion. Artech House Publishers, 1990.

[4] John Robusto James Llinas, & David Noble. "Joint Exploitation Module." In Proceedings of the 1994 Tri-Service Data Fusion Symposium. Applied Research Laboratory. Laurel, Maryland. 1994.

[5] Tom Cool & David Noble. "Intelligence and Object Data Base Generator Tools." In Proceedings of the 1994 Tri-Service Data Fusion Symposium. Applied Research Laboratory. Laurel, Maryland. 1994.

[6] Steve Shaker and Victor Richardson. "Putting The System Back Into Early Warning," Competitive Intelligence Magazine; May-June 2004; pp. 13-17, 2004.

[7] L. Niles, and A. Pease. "Toward a Standard Upper Ontology" in Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds. 2001.