

Structuring Open Source Information to Support Intelligence Analysis

David Noble
Evidence Based Research, Inc.
1595 Spring Hill Road, Suite 250
Vienna, VA 22182, USA
Email address

Keywords: OSINT, Information Extraction and Link Analysis, Search and Retrieval, Fusion, Competitive Intelligence

Abstract

Open source information on the internet can contribute significantly to intelligence assessments. Unfortunately, this information is mostly unstructured text, and varies widely in accuracy, focus, and level of impartiality. Though such unstructured information may be organized for convenient review in information portals, it is difficult to visualize, link, consolidate, summarize, compare, fuse, or otherwise analyze unstructured free text information. Once information is structured, it may be examined and processed using a broad spectrum of powerful commercial tools. This paper reviews the benefits from structuring open source information, describes the criteria that structured information should meet, and outlines a structuring methodology that EBR is not employing for a major U.S. government organization. This methodology employs an advanced service-based architecture for integrating collection and analysis tools, commercial text extraction software, formal ontologies, and tools for operator review and refinement of the machine-structured text.

1. An Example of Structuring

Structuring creates well-defined data records from unstructured material. The following example, based on material that EBR is processing for one of its clients, illustrates the end result of structuring, and serves to illustrate the important properties of structured information.

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.

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 structuring 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

Figure 2. Structured Information

2. Advantages of Structuring

An intelligence analyst has a daunting problem. He must examine potentially enormous amounts of data to extract vital clues. The analyst consolidate and organize the data to make sense of them, and must do so even though essential data may be missing, and even though those pieces that are available may be imprecise and not always trustworthy or accurate. When the data are documents of unstructured text it can be very difficult to put these pieces together, and to see the trends and relationships needed to support effective decision making.

Sarah Taylor (Taylor, 2003) expressed this problem succinctly, stating: “The first difficulty in automatically presenting information from text to the analyst in summarized, consolidated or organized formats is that the original text form itself is not particularly useful. ...If the analyst wishes to quantify his analysis at all, he will somehow have to extract the items of interest from the text and put them in a consistent form. ...The information needs to be in a structured format—discrete items, consistently labeled, or consistent lengths and data types, arrayed in spreadsheets, databases, tables, and the like....”

Structuring the unstructured information, when addressing the issues described in this paper, offers enormous advantages to the intelligence analyst. Once the

text of Figure 1 is transformed into the structured records of Figure 2, the analyst can much more easily examine the totality of information. He can apply powerful visualization tools to help explore and understand the data. He can see trends and discover relationships. He can consolidate the information, making it more succinct and easier to review. He can integrate data from multiple sources using a common format. He can more easily compare the data for conflicting or supporting views, and inspect the data for consistent bias from that source. He can more readily see what is known, what is uncertain, and what is missing. When not familiar with a source, he can compare the information from that source with other sources, and even generate a track record for source credibility (Noble, 2004).

These advantages for intelligence analysis are substantial, but unfortunately until recently structuring could not be feasibly accomplished for most uses. The key to economical and timely structuring is the availability of automated text structuring tools, such as were examined in the Message Understanding Conferences that DARPA sponsored in its Tipster program (Grishman and Sundheim, 1996). These R&D tools of that program were not quite capable enough to support practical text structuring needs. Fortunately, the architectures and tools are now available so that such structuring can be productively accomplished, provided the text extraction tools are supplemented to overcome their current limitations. Section 4 of paper describes how EBR is structuring free text to support analysis, building on the capabilities of a state of the art commercially available text extractor. Before reviewing this methodology, however, the paper reviews some of the issues that structuring should address in order to maximize its value.

3. Desirable Properties of Structured Information

To ensure that the structured information can properly support the needs of the intelligence analyst, the structuring process should address four key issues: 1) quality control for the completeness and precision of the extracted information; 2) use of well-defined concepts within an abstraction hierarchy; 3) management and representation of uncertainty; and 4) management of metadata needed to establish an information audit trail.

3.1 Extraction Quality Control

The structuring process starts with finding the relevant information and extracting it. In our example, it was desired to extract information about company alliances and partnerships. When one occurred in the text, it was hoped that the specific alliance and partnership would be found, and that the extraction information would specify the companies and their roles, the technologies involved, and such specifics about the nature of the association as contract value and place of performance.

In the text showed in Figure 1, the desired information was contained in the five sentences. These are labeled as “events” (the terminology of the text extractor). The label also specifies the location of the extracted material in the broader text.

<p>PWID 1</p> <p>Event 1. Region = 24 – 38</p> <p>Region Indonesia's largest telecommunications company PT Telkom has awarded Ericsson with a contract to provide broadband technology and related services.</p> <p>Event 2. Region = 40 – 55</p> <p>Ericsson will be providing customers in Surabaya, East Java with an Ethernet DSL access solution.</p> <p>Missed event</p> <p>He confirmed the value of the contract to be \$7.5M, with a target start date of late March or early April.</p> <p>PWID 2</p> <p>Event 3. Region 15 – 37.</p> <p>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.</p> <p>Event 4. Region 62 – 75.</p> <p>Ericsson's work for this contract is slated to begin in April of 2005.</p>

Figure 3. Information Extracts

In this particular example, the text extractor found four of the five sentences of interest, missing the one labeled “missed event.” In addition, EBR had not found it cost effective to set up the extraction tool to extract the detailed information about company roles.

These limitations are unavoidable given the current state of the art in text extraction. Because of these limitations, it is important to test and tune the extraction tool to an acceptable and known level of performance. It is also necessary to provide to a facility for operator review and edit.

3.2 Use of Well Defined Terms in an Abstraction Hierarchy

As previously mentioned, the structured records representing the extract information should use well-defined concepts within an abstraction hierarchy. Ontologies provide a way to do this. An ontology provides three services. It creates a logically organized set of concepts organized in an abstraction hierarchy, it defines the properties of each concept using an unambiguous vocabulary, and it can express constraints on these properties.

Each of these properties can be important to the intelligence analyst, but the first two are especially important. The abstraction hierarchy enables reasoning at multiple levels of abstraction. For example, if one wanted to compare extracts about military organizations, one could reason in terms of a “military organization” class. Alter-

natively, if this class were too narrow, then one could just as easily reason in terms of a broader “organization” class include not only military organizations, but political, commercial, educational, and religious ones.

In the structured records, all of the data are labeled in terms of the kind of information they are. Thus, in Figure 2, Ericsson is labeled as an “organization” and “vendor” is labeled as its role. In the ontology used in this example, organization is a text string, and can be anything. In contrast, role is an enumerated variable, and must have only one of a number of previously specified values. These enumerated values help enforce the requirement that the same word will mean the same thing and that different words will mean different things.

The labeling of the data as occurs in an ontology is sometimes regarded as a step in converting data to information (Devlin, 1999). That is, the labeling specifies the kind of thing the data represents, and so is a required step in assigning meaning to the data.

Ensuring consistency with a formal ontology also helps provide interoperability, so that different tools may be more reliably applied and different data sources may be more readily integrated.

3.3 Management of Uncertainty

Information is often incomplete and imprecise, and its accuracy may not be known. The uncertainty representation in the structured record must enable the analyst to know these uncertainties so that he can take them into account in his analysis or decision making. In addition, it must enable the analyst to distinguish between fields that are blank because no information on the needed value was available even though it was sought versus information that was not available but was not searched for.

The example of Figure 2 can reflect some of these uncertainties. For example, the uncertainty of the association start date is represented as a range of possible values, and confidence in the accuracy of this range is represented by the enumerated variable “high.”

The uncertainty representation in Figure 2 is simplified for this paper. The underlying ontology includes a much more complete set of uncertainty representations, including distribution formulas and uncertainty histograms. Confidence in the accuracy of the data often derives from confidence in the source of the data. The structured records store these estimates in the meta-data records.

3.4 Meta data for information audit trail and pedigree

A fourth important property of the structured records is pedigree or audit trail information. Because a structured record is necessarily an abstraction of the source information, there is always a chance that the extraction and structuring process misinterpreted or omitted some of the source information. To be actionable, the analyst must

be able to examine the original unstructured information that the structured records are derived from.

All structured records contain this information. In the example of Figure 2, the audit trail information is accessed through the “communication reports.” Communication reports in the EBR system are structured records that encode the information in a single sentence of the source material. Figure 2 specifies four communication reports because the information in Figure 2 draws on four different records.

Figure 4 shows two of these four reports, each drawn from a different sentence in the source material, and each containing only part of the information in the final consolidated structured record.

Note that the communications records have additional more extensive audit trail information, listed at the bottom of the record in the “pedigree” portion. This information includes the actual sentence that the communication report was built from, a source ID (the published work identifier (PWID)), and the location of the extracted sentence in the source (region). Region is the number of the first and last “token” in the extracted sentence. The text extractor provides the PWID and region. In Figure 5, the two communication records are extracted from consecutive sentences. The processor identifies who or what produced the record. Here, Aerotext™ was the producer. If the material was extracted manually, the producer would be that person’s initials.

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		Surabaya
Providence		East Java
Country		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

Figure 5. Structured Communication Records Representing Information Extracted from Single Sentences

In addition to these audit trail data, the database containing the structured records also include source “pedigree” records, accessed through the PWID number. These records contain information useful for assessing the credibility or trustworthiness of the source. It could contain that source’s accuracy track record for different kinds of information, as EBR suggested in (Noble, 2004). It might also contain information about the, or about possible interests that could generate a rationale for particular self-serving viewpoints.

4. Structuring Methodology

This section describes the architecture and methodology that EBR uses in its to collect and structure records that meet the criteria described above. It relies heavily on a state of the art text extractor, Lockheed Martin’s Aerotext™. It also relies heavily on a formal ontology and on custom-developed methods for refining and consolidating the information that Aerotext™ extracts and structures.

4.1 Overall Processing Flow

Figure 6 summarizes the processing flow in the “War Room” system that EBR uses to extract and structure open source information from the internet (Shaker and Richardson, 2004). In addition to collection and structuring, this system also provides tools to help analysts explore their data to find patterns and relationships and to make forecasts and projections.

The major processing steps are collection (“web harvesting”), text extraction, ontology alignment, manual review and edit, and consolidation. The processing flow generates a sequence of records that successively build the final structured record shown in Figure 2. These records are the Aertotext™ output (Figure 7), the “Communication Record” (Figure 5), and the Consolidated Record (Figure 2).

4.2 Collection and Extraction

Before the source material can be extracted and structured, it must be collected. 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 as blocks of text ready for Aerotext™ processing. The people responsible for information collection identify the primary sources likely to contain the desired information, and then configure Kapow to parse these web pages.

After being collected, the information is ready for extraction and structuring. EBR uses Lockheed Martin’s Aerotext™ to find the relevant information in the collected source materials, to extract the relevant information, and then to structure this extracted information.

Before Aerotext™ can do this, an Aerotext™ specialist must generate a set of 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.

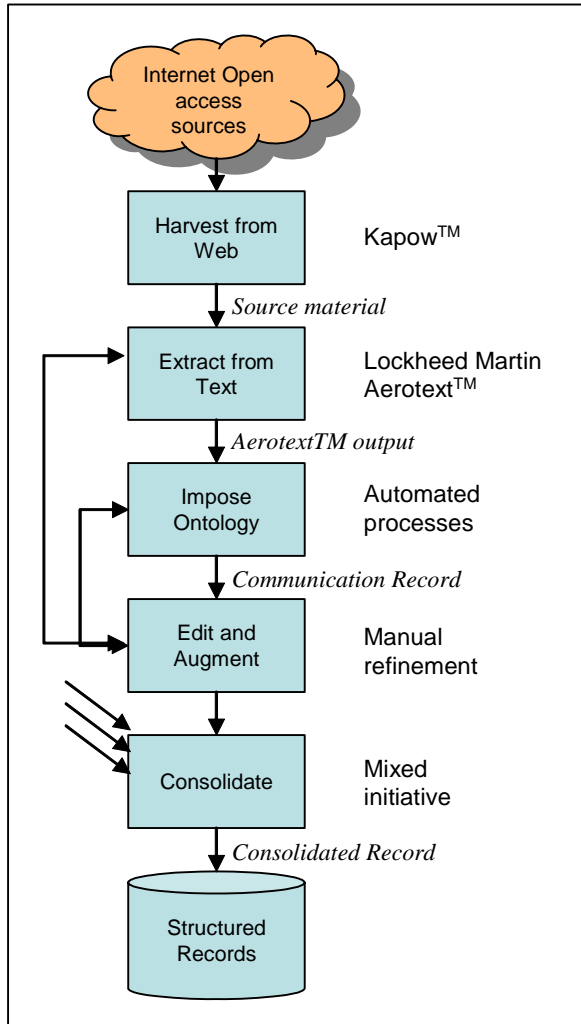


Figure 6. Open Source Collection and Structuring System

The Aerotext rules that we use 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.

Figure 7 shows the Aerotext extraction generated from Events 1 and 2 in Figure 3. Note that it contains the same source pedigree information that the communication records did, except that it doesn't need to note the processor, since this is always Aerotext™ for Aerotext™ output. The out-

put of the Aerotext record is almost the same as in the communication records it feeds, with the exception that the organization in Event 1 is “Telkom” and in the associated communication record is “PT Telkom.” The major difference between the Aerotext output records and the communication records are the additional fields in the communication records for organizational role and for contract date uncertainty.

	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

Figure 7. Aerotext™ output record

4.3 Imposing the Ontology and Editing

The Aerotext™ output records are not constrained by the ontology because Aerotext™ needs to employ a broad vocabulary in order to find the material it should extract. Although the rules that Aerotext™ uses do not depend wholly on recognizing literals, such literals help Aerotext™, and accordingly, it is important not to constrain these. In contrast, the consolidated structured records need to use a well-defined vocabulary for the reasons described in Section 3.2. This processing step aligns the Aerotext™ output with the ontology, transferring the output from the Aerotext-structured records to the ontology-constrained communication records.

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 (Niles and Pease). 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, evidential reasoning.

The record shown in Figure 2 is a simplification of the “Organization Association” class in the ontology. The communication records (Figure 5) are also defined by ontology classes. They serve as a 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.

The EBR system can automatically transfer data from the Aerotext output records into the ontology-defined communication records using a table look up. The system also provides a convenient user interface for manual editing of the communication records, where the operators can align the vocabulary and add information that Aerotext overlooked. In our example, an operator changed "Telkom" to "PT Telkom." More important, the operator was able to fill in the roles for Ericsson and PT Telcom by reading the source material. Finally, the operator by scanning the material in the original source document (the PWID) that was near the extracted text, was able to find a key sentence that AerotextTM had overlooked (the "missed event" in Figure 3). This sentence confirmed the amount of the contract between Ericsson and PT Telkom. Using forms provided by the structuring system, the operator created a new communications record for this sentence.

44 Consolidation

In our use of AerotextTM, the unit for text extraction is a single sentence. That is, Aerotext creates a different output record for every sentence. A communication record is then created for each Aerotext output. Because this sentence-based extraction can produce an unnecessarily large number of records, the EBR system attempts to consolidate all of the communication records that are about the same entity or activity. In this example, it consolidated five communication records, the four derived from Aerotext extraction and the one that the operator created.

To do this consolidation, the system and operators must decide if the communication records are about the same thing. To do this, they first eliminate associations between records with incompatible information (e.g., different organizations) Next, they decide whether to associate the remaining candidates by consider the similarity of the information in the record and the proximity of the source material (since sentences in the same paragraph are often about the same thing).

Once the association decision is made, the communication records can be consolidated. A consolidated record is made from its constituent communication records by populating all of its fields with the data from any of the communication records. When more than one communication record has relevant data, then the consolidation process considers each of these records to create more precise information with reduced uncertainty. In our example, this creates the final structured record of Figure 2.

5. Summary

Structuring unstructured source material can benefit intelligence analysis significantly. By structuring informa-

tion, analysts can determine trends, explore relationships, and sometimes discover the unexpected.

To be most useful, the methodology for structuring data should address four issues. First, it should monitor the collection and structuring process to ensure that these processes are effective. Second, it should enforce use of a standard set of concepts and terms. Third, it should manage and represent uncertainty so that analysts will be aware of the precision of the collected data. Fourth, it should maintain an audit trail and pedigree data so that analysts can review the original source material and assess its credibility and trustworthiness.

EBR is collecting and structuring information in it's war room system. The structuring builds on a state of the art commercial text extractor, Lockheed's AerotextTM. It supplements Aerotext output with software for aligning information with the ontology, for operator review and editing, and for consolidation of multiple data streams to create a more coherent and more useful product to support intelligence analysts.

structured data should Extraction Quality Control Use of Well Defined Terms in an Abstraction Hierarchy Management of Uncertainty Meta data for information audit trail and pedigree

References

- Grishman, R. and Sundheim, B. 1996. Message Understanding Conference – 6: A Brief History. Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, June 1996.
- Devlin, K. 1999. *Infosense: Turning Information Into Knowledge*. W. H. Freeman and Company
- L. Niles, L. and Pease, A. "Toward a Standard Upper Ontology". Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds. 2001.
- Noble, D 2004. Assessing the Reliability of Open Source Information. Proceedings of the 7th International Conference on Information Fusion. Stockholm, Sweden. 28 June – 1 July, 2004. , 2004.
- Shaker, S and Richardson, V. 2004. Putting the System Back Into Early Warning. *Competitive Intelligence Magazine*; May-June 2004; pp. 13-17
- Taylor, S. 2003 Improving Analysis with Information Extraction Technology. Proceedings of the 8th International Command and Control Research Symposium. National Defense University, Washington, D.C., 2003.