Assessing the Reliability 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. However, because the accuracy of this open source information varies widely, the correctness of the information needs to be assessed before it can be used reliably. Current methods for estimating correctness rely on the subjective opinions of knowledgeable people in the field and can vary among evaluators. Today, new data collection and information management tools enable objective reviewer-independent assessment of open source information correctness. These tools support four objective methods for estimating reliability: (1) objective assessment of the historical accuracy of a particular source, by subject matter and viewpoint; (2) self-assessment of reliability from the source itself; (3) consistency of report with prior incidents and with established facts; and (4) consistency of information with other independent reports. This paper describes how these techniques are employed in Evidence Based Research's war rooms to help clients understand the diversity and credibility of viewpoints on client-selected topics.*

Keywords: open source, reliability, objective, fusion

1 The assessment problem

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.

Though potentially of great value, it is often difficult to take advantage fully of open source information. The difficulty in doing so has many causes: it is not always easy to find some key information because only a small fraction of the Internet is indexed; it is not easy to extract and combine key information because most of the reports are free text; and it is not easy to assess report credibility for reasons listed below. This paper focuses on the last of these obstacles, assessing report credibility. The methods for assessing credibility also help overcome the other obstacles to using open source information.

The ability for anyone to post information on the Internet and the lack of any regulation over content accuracy fosters a great deal of erroneous data and misinformation, which is intermingled with high valued nuggets. Thus to confidently use open source information, its accuracy needs to be assessed. Unfortunately, this assessment can be difficult. Current methods of evaluating open source information (e.g., [1]) are subjective and can require considerable user expertise. For example, most of the criteria in Robert Harris' CARS (credibility, accuracy, reasonableness, support) checklist require expert judgment on such issues as balance of the article, believability of information, and political bias of source.

Because these assessments depend on the skill of human analysts, the resulting analyses based on these reports are proportional to the level of knowledge and the experience of the people making these assessments. This dependency can lead to uneven results. For example, the recently released Congressional report of the Joint Inquiry into the Terrorist Attack of September 11, 2001 [2] amplifies this by stating that "the quality of counter-terrorism analysis was inconsistent, and many analysts were inexperienced, unqualified, under-trained, and without access to critical information. As a result, there was a dearth of creative, aggressive analysis and persistent inability to comprehend the collective significance of individual pieces of intelligence."

2 Approach

This paper describes how to supplement these subjective assessments with more objective ones. In this approach, individual reports are not assessed in isolation, but rather in the context of other previous reports on similar topics from that same source, of current reports on the same topic from other sources, and of other known facts and precedents.

The assessment methodology considers the following three considerations when assessing the reliability of information reported by a particular source:

1. The actual historical reliability of that source on similar events or subjects, taking into account the report's self-assessment of reliability.
2. The report's consistency with confirmed facts and precedents.
3. Its consistency with information available from other sources.

Evaluating consistency with other sources requires that that information be fused in order to create the needed basis for comparison. Figure 1 outlines a generic fusion process for collecting, structuring, and fusing open source information.

This process begins with the collection of unstructured open source information, such as the information shown in Figure 2. This information is then structured into event records whose attributes are formally defined in an ontology. To do this, each source type needs a source-specific "translator." Figure 1 shows translators for imagery and voice as well as free text to show the design scalability. We address only free text in this paper. After initial structuring, these event records are then conditioned by the actual historical reliability of that source on similar events or subjects, and the report's consistency with confirmed facts and precedents. The fusion process then combines these records, drawing on all available current and historical information to estimate event uncertainties.

In the past, this process would not have been practical, for it would have been infeasible to collect, structure, and analyze the number of related reports required for an objective assessment. It requires, for example, collecting reports from multiple sources on the same incident, multiple reports from the same source on similar incidents, background facts, and reports on precedents. Today, however, the technology exists to accomplish the collection and analyses needed to support an objective assessment of information reliability. Figure 3 depicts the major components of the system that EBR uses to collect and structure open source information. At EBR we use primarily Intellisonar™ to collect and process open source information, converting Web pages into labeled blocks of 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.

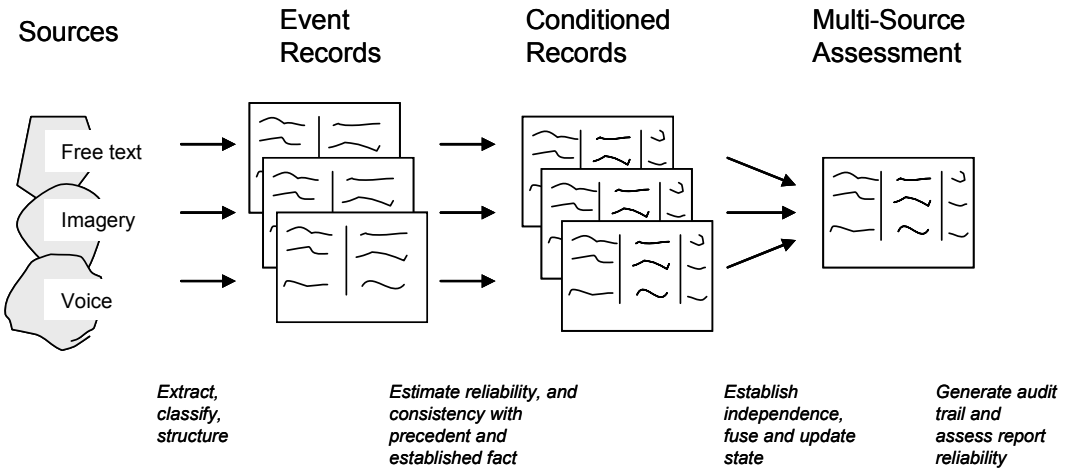

Fig. 1. Generic Fusion Process

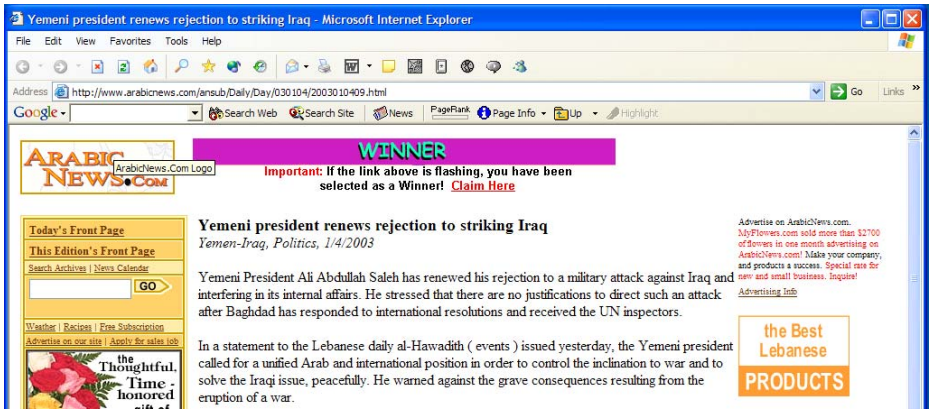

Fig. 2. Open Source Information

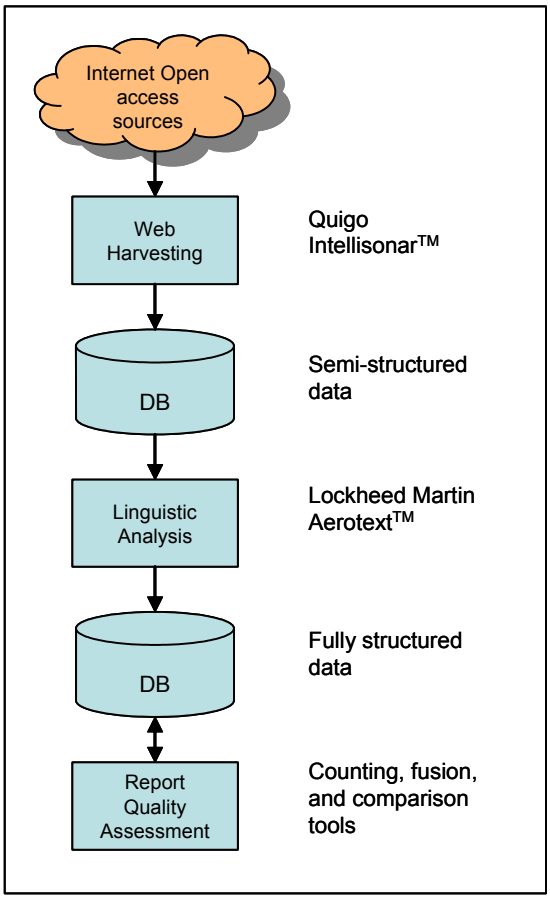

Fig. 3. Open source collection, structuring, and analysis tools

3 Assessing the historical reliability of a source

This assessment provides for the objective measurement (by counting) of source reliability on various topics. It is not directly concerned with the reasons for possible low reliability, such as self-interest or ideology, but instead objectively measures what the actual reliability has been for previous reports of this type. Note that the reliability assessment is performed individually for each element of a report (e.g., in a terrorist event, the reliability assessment for the "type of event" field is estimated separately from the reliability assessment of the "performing organization" field).

The actual historical reliability of a source is estimated by counting the number of times a report says that something has some particular value when in fact another value was correct. The specific form of this reliability estimate is the probability that a source reports something as "y" when its actual value was "x." These reliability computations need to be conditioned on the report's self-assessment of its own reliability. In fact, this historical reliability is actually an objective measurement of the accuracy of the report's self-assessment. Further, because the reliability is expected to vary depending on the subject reported on and the viewpoint expressed, the reliability needs to be further qualified by subject and viewpoint.

This historical reliability estimate requires a way to gather and structure the historical reports, a way to determine whether the report is correct or not, and a way to classify the report's self-assessment and domain topic.

3.1 Gathering and structuring the reports

Currently available COTS tools, many of which did not exist even 2 years ago, are now available to improve the efficiency of information collection and structuring enormously [3]. As summarized in Figure 3, the basic approach (1) collects huge amounts of information from diverse open sources, (2) creates a preliminary structure to the information, and then (3) performs additional text processing to tag the information with the appropriate markers for analysis.

The collection and parsing tools "harvest" unstructured data and parse the various pieces into labeled text blocks, usually using an XML tagging system. These labeled text blocks might include <TITLE>, <AUTHOR>, <BYLINE>, <DATE>, <SOURCE>, <BODY>, etc. This parsing is the first step in applying structure to unstructured text. The next step is to structure the text in the body. In this step, the free text processor extracts and deposits information into a domain-specific template. For example, a terrorism template would have slots for the type of an attack, its time and place, the terrorist organization, the damage, and casualties.

The structuring processes also need to extract the report's self-assessment of reliability. Examples of such self-assessments can be found in phrases such as:

"…30 people witnessed a man dropping a briefcase immediately prior to the explosion…"

"…terrorism experts have concluded that Al Qaeda was responsible for…"

"…rumors have suggested that Osama bin Laden is behind the recent attacks…."

Such phrases provide important information about the uncertainty of the reported information.

In order to apply these assessments to the overall assessment, the text processor must interpret these phrases as numerical probabilities. Though such probabilities can only be approximate, such rough estimates are adequate for supporting the reliability estimates.

3.2 Determining correctness of historical reports

The correctness of reports often cannot be determined at the time of the report, but must be determined after the "truth" has a chance to emerge. Because the reports being used to objectively quantify a source's historical reliability are drawn from the past, often enough time has passed for the truth of the incident to have been established.

This truth can be determined by fusing the most recent reports on the incident using the methods to be described later. If the sources contributing to the fusion product have high self-assessments of their reliability and if these sources agree and are independent, then this fusion product can provide a standard for determine a source's historical correctness.

3.3 Classifying reliability and domain topic

Because the objective assessment of source reliability requires many different reports on the same kind of incident, the assessment requires that incidents be classified abstractly. The needed levels of abstraction are defined using an ontology.

The required abstraction hierarchy defines the concepts needed to structure the information in open source reports. These definitions must be understandable by automated text extraction engines and by knowledge management systems. This abstraction hierarchy addresses reliability and uncertainty, source types, and different key characteristics of relevant domains. For example, the domain of terrorism includes kinds of terrorist activities, methods, and organizations. The abstraction hierarchy enables the reliability assessment and fusion methods to reason at multiple levels of abstraction. EBR is building its 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.

Table 1 is an example of part of a possible abstraction hierarchy for terrorist events, sources, and reliability evaluations.

Table 1. Example of an Abstraction Hierarchy

Terror Events	Terrorism Actions	Bombing Kidnapping Execution
	Activities supporting terrorism	Fund raising Recruiting Advocating
	Counterterrorism	Military action Law enforcement action
Entities	People	Perpetrators Victims Supporters
	Places	Country City
	Time	Date Time
Source	Newspaper or News Organization Columnist Political Figure/Administration Official Academic Tabloid Internet News Source	
Source Evaluation of Validity	Eye Witness Confirmation Inferences from Eye Witness Testimony Analysis (meets the pattern of)	

4 Evaluating a report's consistency with precedents and confirmed facts

This method of determining source reliability builds a database of instances about entities and activities. A precedent for an attribute or relationship having a particular value is the existence within this database of an attribute or relationship with that value. The strength of a precedent can be determined using the same technique as was used to quantify the historical reliability of a source: by counting how often the particular variable or relationship values have been observed in the past. As in that case, the variables or relationships need to be defined at a level abstract enough so that a sufficient number of examples can be found. This is another important application of the abstraction hierarchy and concept ontology.

Counting the fraction of times variable or relationship values have been previously observed is a conventional way of establishing prior probabilities. For example, the prior probability that a terrorist organization would carry out an attack in a particular way would be the relative frequency that that organization attacked in that manner. Using a prior probability calculated this way to estimate a posterior probability from its prior can unfortunately produce serious error, for just because groups have never done something in a particular way before does not mean they will not in the future. Consequently, this notion of "prior" needs to be interpreted as "consistent with capability," with capability inferred from precedents.

Determining a report's consistency with confirmed facts entails identifying report elements that can be independently checked against established facts, such as the names of people serving as government officials. When conflicts are found, then the a priori probability of the statement can be set very low, depending on the credibility of the "established" fact.

Both determining precedent and checking consistency with confirmed facts are important to consider. In the case of the reported plan for Iraq to import uranium from Nigeria, there was ample precedent to substantiate a high prior probability for the report. However, because the report was inconsistent with supplemental information (for the person signing the report was not in the Nigerian government at the time of the report), the report was shown to have very low credibility.

5 Determining consistency of information available from other sources

The third way to estimate report reliability is to compare the consistency of the information whose reliability is to be assessed with other information on the same subject. Within our framework, this is accomplished by comparing the structured report representing the information to be assessed with the product generated by fusing the structured reports obtained from other information sources. By weighing and combining all relevant information from multiple sources the fusion product provides the best summary of that information. It also automatically takes into account the extent to which different reports support or contradict one another.

Because this fusion product specifies the mean and uncertainties of all variables in a structured report, it is easy to determine the consistency of the information to be assessed (the target information) with information from other sources (the fused information). If all the uncertainty

ranges of target information overlap the uncertainty ranges of the fused information, the target information is consistent with the other information. If the uncertainty ranges of the target information do not overlap the ranges of the fused information, the target information is inconsistent with the other information.

Fusing the information from multiple open sources entails the following steps:

1. Collect the information to be fused from multiple sources.

2. Check for each source that the information is actually about the same subject as the report to be assessed.

3. Structure the information from each source using an appropriate template.

4. Determine the extent to which the sources are independent and the extent to which they are relaying information from a common source.

5. Identify the uncertain report attributes (e.g., those for which the sources differ).

6. For each of these attributes, identify the range of reported attribute values.

7. Estimate the reliability for every source and every attribute, given each source's historical track record and self assessment of reliability.

8. Based on an analysis of precedents and consistency with established fact, estimate a prior probability for each attribute value.

9. Considering the dependencies among reports, estimate the joint probabilities for multiple sources.

10. Update the state estimates for the information being fused.

As discussed earlier, the first three of these steps draw on the open source data collection and structuring processes that are now possible using advanced collection and knowledge management tools. This section discusses the remaining steps under the topics: (1) determine dependencies among reports; (2) condition report attributes; (3) associate reports and manage hypotheses; and (4) fuse and refine state estimate.

5.1 Determine dependencies among reports

The number and credibility of independent confirming and contradicting reports is a powerful means for assessing the correctness of reports. Additional independent reports can contribute substantially to the assessment of the reliability of a single information source. However, multiple reports that merely repeat material from a common source contribute no additional information beyond that provided by their common source. Determining report independence is therefore essential for handling multiple reports on the same topic.

The collection and structuring methods described previously in this paper help to identify when two reports are restatements or paraphrases of the same source. Retransmission is commonly the case when news sources publish articles from the various newswires such as Reuters and API. The news is simultaneously transmitted by various distribution channels and would appear to be separate reports but in fact is the same article from a common source.

Methods that help determine whether information is original or draws on sources used by other reports includes checking for explicit references to an original source, checking for similarity between title, author, and byline, or checking for the time and date the original source posted the information. It is also possible to infer a common source when two reports use identical language.

5.2 Condition report attributes

The text extraction and structuring step creates a table for each information report. For reports on incidents, for example, the table specifies general attributes of an incident, such as the type, time, and place of the incident and the responsible organization. Information in this form cannot be fused until it is "conditioned," e.g. until attribute uncertainty and reports dependencies are established. The processes described in Sections 3 and 4 can provide this conditioning.

The methods for assessing the historical reliability of a source, given the domain and source's self-assessment can provide for each attribute a probability that that attribute has the value reported. The methods for estimating the consistency with established facts and with precedent can further refine the estimate of these priors. Note that in order to support fusion, uncertainty must be estimated separately for each of the report fields rather than for the report as a whole. This needs to be done because in any report some of the attributes might be highly reliable while others may be speculative.

The uncertainty and independence estimates are essential for fusion of information. It is not possible to estimate the likelihood that two reports are referring to the same thing without these uncertainty estimates. The estimates are also required to balance the credibility of alternative accounts of an incident. And of course it is impossible to estimate the uncertainty of the fused product without the uncertainty estimate of the contributing reports.

5.3 Associate reports and manage hypotheses

Deciding whether reports are actually reporting on the same incident is the report association problem, which for many fusion problems is the principal challenge. The key to report association is the uncertainty specifications for each attribute. It is not possible to associate reports without these uncertainties. 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.

5.4 Fuse and refine state estimate

Once it is determined that two reports reference the same entity or event, the information in these reports can be combined to refine the state estimate. Practical methods for doing this can be based on Bayesian reasoning, but will also draw on heuristics to ensure the scalability of the algorithm in complex environments.

The basic Bayesian expression for combining statements from two sources on the value of an incident attribute is:

Prob(attribute is x | source "A" says it is y and source "B" says it is z) =

[Prob (source "A" says it is y and source "B" says it is z | attribute is actually x) * Prior Prob (attribute is x)] /

[Prob(source "A" says it is y and source "B" says it is z)]

The a priori probability that source A would report that the attribute has each particular value given that it has that or any other particular value is estimated using the techniques described in Section 3, objectively quantifying the historical tendencies of various sources to report various conclusions. For example, a news source may choose to attribute all terrorist attacks to Al Qaeda, no matter who is actually responsible. Over time, data on who actually performed these attacks will become available. If historically a source always attributes attacks to Al Qaeda even when Al Qaeda is not responsible, then a subsequent report from that source that Al Qaeda is the terrorist agent would not provide any useful information on the attacker's identity. The Bayesian equation reflects this history, and the formula automatically ignores the source. In this case, the posterior probability is the same as the a priori probability.

Estimating the joint probabilities for multiple sources having various combinations of reports is essential to avoid inadvertent double counting. These estimates can be approximated by assuming either that the reports are the same or that they are completely independent. If they are thought to be the same, as judged using the reasoning described in Section 5.1, then the additional reports can be discarded. If it is acceptable to approximate them as independent, the expression Prob (source "A" says it is y and source "B" says it is z | attribute is actually x) can be replaced by the product Prob (source "A" says it is y | attribute is actually x)*Prob (source "B" says it is z | attribute is actually x).

6 Generate an assessment audit trail

An assessment audit trail explains the basis for the reliability and uncertainty assessments. For example, it can list the reasons for and against particular conclusions and can point to the reports responsible for these reasons. These audit trails are an essential part of the assessment product. They are needed in order to judge the trustworthiness of the assessment and for explaining the assessment to others.

Because the internal structure of fusion hypotheses and their links to supporting data can be complex, a more easily understandable audit trail contributes significantly to the usefulness of the results. This audit trail helps people quickly understand the evidence for and against alternative interpretations of the data. It also helps them quickly assess the reliability the open source information, helps them understand the basis of the fusion conclusions, and helps them integrate the open source information with classified data.

A format used in previous Level III fusion work [4,5] qualitatively summarized the arguments for and against each hypothesis using the categories of information shown in Table 2.

Table 2: Summarization of Arguments For and Against a Fusion Hypothesis

Arguments supporting a hypothesis	**Arguments opposing a hypothesis**
• Confirming evidence	• Conflicting evidence
• Lack of alternative hypotheses able to explain the evidence	• Alternative hypotheses able to explain the data
	• Available data that the hypothesis cannot explain
	• Data expected if the hypothesis is true, but that was not obtained

Given this understanding of the fusion product, it is easy to understand the reasons for an assessment of a report's reliability:

1. The report's self-assessment is stated within the report itself; e.g. the report explicitly states that the organization responsible for an event is not confirmed.

2. The historical record of a source's accuracy on a particular subject is documented by listing the number of times in the past that that source was accurate on that subject.

3. Similarly, consistency with precedent and established fact can be easily understood since all precedents are documented and conflicts with established fact flagged.

4. Consistency with other information can be determined easily by comparing the uncertainty ranges of the information being evaluated with the uncertainty ranges of the fusion product generated from other sources. The credibility of the fusion product can in turn be judged from the information in Table 2.

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 was hard to do because of the difficulty of finding key information, extracting and structuring free text, and evaluating source credibility. 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 objectively evaluate source credibility, and so combine multiple sources into information landscapes to support superior situation assessments and decisionmaking.

References

[1] Robert Harris. *A Guidebook to the Web*. Dushkin/McGraw Hill. 2000.

[2] Report of the Joint Inquiry into the Terrorist Attacks of September 11, 2001. The House Permanent Select Committee on Intelligence and the Senate Select Committee on Intelligence. July 23, 2003.

[3] Steve Shaker. "Connecting the Dots: War Room Team-Based Analysis." InterSymp-2003 Proceedings of the focus symposium on Collaborative Decision-Support Systems, July 2003.

[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.