RESEARCH ARTICLE

International Journal of Multidisciplinary and Scientific Emerging Research

Available at http://www.ijmser.com/ (ISSN 2349 – 6037)

Open Source Intelligence of Web Mining

Raghavendran V

Department of Information Technology,
University /College Higher College of Technology, Oman

Accepted 26 September 2014, Available online **13** October 2014, **Vol.3, No.1 (October 2014)**

Abstract

Open source intelligence of web mining is the extraction and analysis of directly or indirectly, publicly available information from various source using Information extraction techniques. OSINTWM operations support other intelligence, surveillance, and reconnaissance (ISR) efforts by providing foundational information that enhances collection and production. As part of a multidiscipline intelligence effort, the use and integration of OSINTWM ensures decision makers have the benefit of all available information. Two important terms in these complementary definitions are — Open Source, which is any person or group that provides information without the expectation of privacy — the information, the relationship, or both is not protected against public disclosure. Publicly Available Information, which is data, facts, instructions, or other material published or broadcast for general public consumption; available on request to a member of the general public: lawfully seen or heard by any casual observer; or made available at a meeting open to the general public Collecting webpage content and links can provide useful information about relationships between individuals and organizations. Properly focused, collecting and processing publicly available information from Internet sites can support understanding of the operational environment.

Keywords: *open source, search engine,*

1. INTRODUCTION

Internet Search Techniques –

STEP	TECHNIQUES AND PROCEDURES
Plan Search	• Determine operations and computer security risks and protective measures. • Use mission and specific information requirements to determine objective and search terms. • Write all search terms down. • Collaborate with librarians and other analysts to determine potential information sources. • Select the search tools and sources that will best satisfy the objective. (These may be on classified systems vice the Internet.)
Conduct Search	• Use approved hardware and software applications. • Use authorized government or commercial Internet service provider. • Search only for information for which the organization has an authorized and assigned mission in accordance with AR 381-10. • Based on requirements, software, and tools of the chosen search engine or resource, conduct search using methods such as keyword searching, field searching, or Natural Language techniques.
Refine Search	• Browse or scan results for relevancy, pertinence, associated terms, discovery of new concepts and terms to follow up on, and irrelevant terms to exclude in more refine searches. • Compare the relevancy of the results to objective and indicators. • Compare the accuracy of the results to search parameters (keywords, phrase, date or date range, language, format, etc). • Compare the results from different search engines to identify missing or incomplete information (for example, one engine's results include news articles but another engine does not).
	• Modify the keywords. • Search within results. • Search by field. • Search cached and archived pages. • Truncate uniform resource locator.
Record Results	• Record relevant source information—as a minimum, URL (location), date accessed, name and date of file of document title, and author or organizations. • Save content. • Download files. • Identify Intellectual Property.

Table 1: Searching Techniques and Procedure

The ability to search the Internet is an essential skill for open source research and collection personnel. The Internet provides access to web pages and databases that hold a wide range of information on current, planned, and potential operational environments. Techniques and procedures for searching the Internet.

1.1 Cyber Security

The Internet is not a benign environment. There are operations and computer security risks to searching the Internet and interacting with Internet sites. Searching the Internet can compromise OPSEC by leaving "footprints" on visited sites. Visiting Internet sites can compromise computer security by exposing vulnerability or providing information that exposing the computer and the network to malicious software or unauthorized access. Users must be vigilant to potential threats; use only authorized hardware and software; and comply with OPSEC measures.

Awareness of what information the user's computer provides to each server and site on the Internet is the beginning of effective cyber security. Just by visiting a site, the computer transmits machine specifications such as operating system and type and version of each enabled

software program, security levels, a history of sites visited during that session, cookie information, user preferences, communication protocol information such as an IP address (for the user and hosting or proxy server), enabled languages, and other computer profile information such as date and time (and time zone), referring URLs (the previous site visited), and more. Available on unprotected computers could be the email address of the user, login information, their certifications. In addition to computer vulnerabilities, just knowing where the research comes from may affect the page accessed. Sites frequently redirect visitors to alternate web pages (or totally block access) based on what user is searching for, where the user is located, what language the user is searching in, and what time of day the user accessed the site.

Uniform resource locator information from the previous sited visited (referring URL) is frequently an OPSEC issue. It identifies some characteristics and interests of the user to the visited site, server, and country. While necessary for an effective search, the use of specific, focused, search terms such as locations, names, and equipment have obvious OPSEC implications.

1.2 Plan Search

Intelligence personnel use their understanding of the supported unit's mission, the SIRs, indicators, and the Internet to plan, prepare, and execute their search. The SIR helps to determine what information to search for and where to look. The SIR provides the focus and initial keywords those intelligence personnel

1.3 Conduct Search

Intelligence personnel should avoid the temptation of using one favorite search engine to the exclusion of others. Each search engine has its strengths and weaknesses. Organizational standards, research experience, and peer recommendations guide the selection of which search engine to use in any particular Situation. Generally, a thorough search often requires the use more than one search engine and even then, the information may not be complete. As a rule, if a trained analyst or collector cannot find the information using multiple search engines and common search techniques within 30 minutes, it is possible that the information is not on the Internet, not indexed, or not in a retrievable format. At that point, the analyst or collector should seek assistance from other personnel, digital but non-Internet resources such as commercial and in-house databases, and non-digital resources available at government or university libraries.

Search Engines

Intelligence personnel use search engines and search terms to locate Internet sites and find information within the Internet site. Search engines allow the user to search for text and images in millions of web pages. The different commercial and government search engines vary in what they search, how they search, and how they display results. Most search engines use programs called web crawlers to build indexed databases. A web crawler searches Internet sites and files and saves the results in a database. The search engine, therefore, is actually searching an indexed database not the content of the site or an online database. The search results also vary between search engines because each engine uses different web crawlers and searches different sites. Most engines display search results in order of relevancy with a brief description and a hyperlink to the referenced Internet file or site.

Web Crawler

Search engines have an index database built by a web crawler. The web crawler or spider is a different application than the search engine. The crawler is like some voracious monster with an insatiable appetite, it roams the Internet 24 hours a day, 7 days a week, searching for information. Once it finds a Website, it then indexes and saves it in a database relevant to the search engine. Some search engines have their own spiders while others use Commercial contracted spider programs to develop their databases. In addition, each spider may use a different approach to acquiring data. One spider may be programmed to research only the titles of web pages and the first few lines of text. Other spiders research virtually the entire website with the exception of graphics or video files. Because search engines may use different web crawler software with different ways to index and save data, each separate search engine may yield different results. Also, the search engine provider can supplement or alter the spider software's index to ensure the website of specific customers appears in the index.

Relevancy Formulas

The relevancy formula evaluates how well the query results match the request. For web pages that are commercially oriented, designing the page to achieve the highest ranking has become an art form. For some search engines, the process is simple, the higher the bid, the higher the site's ranking. Search engines are continually changing their relevancy Formulas in order to try to stay ahead of web developers. Some web designers, however, load their sites with words like "free," "money" or "sex" in an attempt to influence the search Engine's relevancy formula. Other web designers engage in practices called "spamdexing" or "spoofing" in an attempt to trick the search engine. The significance of the relevancy formulas to the user is the importance of understanding that the keyword in the search does not necessarily yield the same results with every search engine. This becomes obvious when the user considers that relevancy formulas vary from search engine to search engine and are in a constant state of evolution. In some formulas, the placement of the keywords yields different results if rearranged because the search engine's relevancy formula places more emphasis on the first words in the search string. Relevancy formulas may also assume importance depending on the type of search being done. For instance, a field-search, which is

limited to the webpage itself (for example, title, URL, and date), may be more critical than a full-text search.

As search engines evolve, some engines have become adept at finding specific types of information such as statistical, financial, and news more effectively than other engines. To overcome this specialization, software engineers developed the Meta search engine. The Meta search engine allows the user to query more than one search engine at a time. On the surface this would seem to be the final answer to the search question; just query all search engines at one time. Unfortunately, it is not quite that easy.

Since it must be designed to work with all search engines that it queries, the Meta search engine must strip out each search parameter to the lowest common denominator of each search engine. For example, if a particular search engine cannot accommodate phrases in quotation marks or a type of Boolean function then the Meta search engine will eliminate that function from the search. The resulting search, in many Instances, then becomes too broad and less useful than a well-formatted search using a search engine that the user is familiar with and that is known to be good at locating the type of information required.

With an understanding how search engines work, intelligence personnel—

- Conduct an initial search using unique key words or key word combinations and, if possible, multiple search engines. Avoid using one search engine to the exclusion of others.
- Evaluate the relevance and accuracy of the search results to research objectives, indicators, and search parameters. Do not rely on the relevancy formula of the search engine, particularly commercial search engines, to list the most relevant information source at the top of the list.
- Conduct follow-on searches using refined terms and methods. Refining terms includes inverting the word order, changing the case, searching common misspellings, correcting spelling, and adjusting search terms. Refining search methods includes searching within results that are similar to the desired information.

Search by Keyword

In keyword-based searches, the intelligence personnel should consider what keywords are unique to the information being sought. The analyst or collector needs to determine enough keywords to yield relevant results but not so many as to overwhelm them with a mixture of relevant and irrelevant information. They should also avoid common words such as "a," "an," "and," and "the" unless these words are part of the title of a book or article. Most search engines ignore common words. For example, if looking for information about Russian and Chinese tank sales to Iraq, the analyst or collector should not use tank as the only keyword in the search. Instead, they should use additional defining words such as "**Russian Chinese tank sales Iraq.**"

In some search engines, Boolean and Math logic operators help the analyst or collector establish relationships between keywords that improve the search. (For example, **Russian tank).** If they want to exclude Chinese tank sales from the search result then he uses **(Russian tank) NOT (Chinese tank) sale Iraq** in the search. The analyst or collector can also use a "NEAR" search when the relationship and the distance between the terms are well established. For example, if the analyst or collector is looking for incidents of earthquakes in Pakistan and news articles normally place the place name of the location of an attack within five words of "earthquake" in the title of body of the article then they use earthquake NEAR/5 Pakistan in the search.

FUNCTION	BOOLEAN	EXAMPLE
Must be present*	AND	Chemical AND weapon chemical +weapon
Must not be present	NOT	Africa NOT Sudan Africa -Sudan
May be present	OR	chemical OR biological
Complete phrase	" "	"Chinese tank sales to Iraq"
Nested	( )	(Shining Path)
Near**	NEAR	"White House" NEAR "airspace incursion"
Wildcards	word* or *word	gun* (gunpowder, gunsight)
Stopwords***	"" ""	""OR"" (do not ignore OR)

Table 2: Boolean and Math Logic Operators.

Search in Natural-Language

An alternative to using a keyword search is the natural language question format. Most of the major search engines allow this capability. The analyst or collector obtains the best results when the question contains good keywords. One of the major downsides to this technique is the large number of results. If the needed information is not found in the first few pages then they should initiate a new search using different parameters.

1.4 Refine Search

Normally, the first few pages of search results are the most relevant. Based on these pages, the analyst or collector evaluates the initial and follow-on search to determine if the results satisfied the objective or requires additional searches. During evaluation, they compare—

- Relevancy of the results to the objective and indicators.
- Accuracy of the results to search parameters (keywords, phrase, date or date range, language, format).

- Results from different search engines to identify missing or incomplete information (for example, one engine's results include news articles but another engine does not).

Modify the Keyword

If initial search attempts are unsatisfactory, the analyst or collector can refine the search by changing the following:

- **Order.** Search engines may place a higher value or more weight on the first word or words in a multiple word or phrase search string. Changing the word order from "insurgents Iraq" to "Iraq insurgents" may yield different search results.
- **Spelling/Grammar.** Search engines attempt to match the exact spelling of the words in the search string. There are search engines that do recognize alternate spellings or prompt the user to correct common misspellings. Changing the spelling of a word from the American-English "**center**" to the British-English "**centre**" may yield different results. Changing the spelling of a transliterated name from "**Al-Qaeda**" to "**al-Qaida**," "**al-Qa'ida**," "**el-Qaida**," or "**al Qaeda**" generates different results that may be useful depending upon the objective of the search. Some search engines provide this capability for a "sounds like-type" search that eliminates or reduces the manual entry of each variation. Looking for common misspellings or common, grammatically incorrect, short phrases may be useful in yielding results from a source for which English is a second language or the language of the webpage is in a second language for the web designer or web contributor.
- **Case**. Search engines may or may not support case sensitive searches. Like spelling, some engines attempt to match the word exactly as entered in the search. The intelligence personnel should use all lowercase letters for most searches. When looking for a person's name, a Geographical location, a title, or other normally capitalized words then the intelligence personnel should use a case sensitive search engine. Changing the case of a word from "**java**" to "**JAVA**" changes the search result from sites about coffee to sites about a software program.
- **Variants.** Intelligence personnel use terms that are common to their language, culture, or geographic area. Using variants of the keyword such as changing "policeman" to "**cop**,""**bobby**," "**gendarme**," **carabiniere**," "**policía**," "**politzei**," or other form may improve search results.

Search within Results

If the initial or follow-on search produces good but still unsatisfactory results, the analyst or collector can search within these results to drill down to the web pages that have a higher probability of matching the search string and containing the desired information. Most of the popular search engines make this easy by displaying an option such as "search within these results" or "similar pages" that the user can select. Selecting the option takes the analyst or collector to web pages with additional, related information.

Search by Field

In a field search, the analyst or collector looks for the keywords within the URL as opposed to searching the entire Internet. The best time to use a field search is when the search engine returned a large number of web pages. While capabilities vary by search engine, some of the common field search operators are—

- **Anchor:** Searches for webpage's with a specified hyperlink.
- **Domain:** Searches for specific domains Like: Searches for webpage's similar or related in some way to specified URL.
- **Link:** Searches for specific hyperlink embedded in a webpage.
- **Text:** Searches for specific text in the body of the webpage.
- **URL:** Searches for specific text in complete Web addresses.

With the millions of URLs on the web, the analyst or collector is faced with a myriad of sites that may or may not actually be produced and maintained by the type of organization represented by the majority of web pages in that domain (see Table.3). Certain domains, such as ".mil," ".edu," and ".gov"are consistently reliable as being administered and authored by those organizations. Several domains have, over the years of ever-increasing numbers of Internet participants, become highly suspect as to the validity of the organization using such a domain extension. In particular, the open source information gatherer must not take ".org," ".info," or ".net" extensions as necessarily produced by a bona fide organization for that domain. Each country has a two-digit digraph and registers domains with Internet Assigned Number Authority (http://www.iana.org). The country digraph is important because it indicates the site in another country. For example, .uk (United Kingdom) has more than a billion sites indexed; .cn (China) has 700 million sites; .fr (France) has more than 600 million sites, while the domains you listed have fewer than a million sites each (for instance, .aero, .jobs, .museum, .pro).

Search in Cache and Archive

Sometimes a search or an attempt to search with results returns a URL that matches exactly the search objective but when the analyst or collector tries to link to the site, the link or the site is no longer active. If the search engine captures data as well as the URL locator, they can select "cached" link to access the original data. Another technique is search in an Internet archive site such as www.archive.org for the content. The analyst or collector needs to be aware that this information is historical and not subject to update by the original creators.

DOMAIN	DESCRIPTION	OPERATOR/SPONSOR
.Aero	Reserved for members of the air-trasport industry	Société Internationale de Télécommunications Aéronautiques
.biz	Restricted to business	NeuLevel, Incorporated
.com	Unrestricted top-level domain intended for commercial content	VeriSign Global Registry Services
.coop	Reserved for cooperative associations	Dot Cooperation limited Liability Company
.edu	Reserved for postsecondary institutions accredited by an agency on the US Department of Education's list of Nationally Recognized Accrediting Agencies	Educause
.gov	Reserved exclusively for the US Government	US General Services Administration
.info	Unrestricted top-level domain	Afilias Limited
.int	Used only for registering organizations established by international treaties between governments	Internet Assigned Number Authority
.jobs	Reserved for human resource managers	Dot Cooperational Limited Liability Company
.mil	Reserved exclusively for the US military	US Department of Defense Network Information Center
.museum	Reserved for museums	Museum Domain Management Association
.name	Reserved for individuals	Global Name Registry
.net		VeriSign Global Registry Services
.org	Intended for noncommercial use but open to all communities	Public Interest Registry
.pro	Restricted to credentialed professionals and related entities	RegistryPro

Table 3: Domain Detail.

1. 5 Record Results

Intelligence personnel must save the search results that satisfy the research objective. Saving the results enables the analyst or collector to locate the information later as well as to properly cite the source of the information in intelligence reports and databases. While printing a hardcopy is an option, a softcopy (electronic) record of the search results provides a more portable and versatile record. Also, some intelligence organizations have software tools specifically designed for creating a complete record of the webpage content and metadata. The following are some basic techniques for saving an electronic record of the search results.

- **Bookmark.** Bookmark the link to the webpage using the "bookmarks" or "favorites" option on the Internet browser.
- **Save Content.** Save all or a portion of the webpage content by copying and pasting the information in text document or other electronic format such as a field within a database form. The naming convention for the softcopy record should be consistent with unit electronic file management standards. As a minimum, the record should include the URL and retrieval date within the file.
- **Download Files.** Download audio, image, text, video, and other files to the workstation. The naming convention for the softcopy record should be consistent with unit electronic file management standards.
- **Save Webpage.** Save the webpage as .mht, .pdf, .doc, .html—or other specified format—that creates a complete, stable record of the webpage content. It may be necessary to include the date and time in the file name in order to ensure a complete citation for the information.
- **Record Source.** As a minimum, record the author or organization, title, publication or posting date, retrieval date, and URL locator of the information in a citation format that is consistent with the American Psychological Association and Modern Language Association style manuals.

The following is an example of a American Psychological Association citation for an Internet document:

BBC News (2005). Sudan: A Nation Divided. Retrieved 16 May 2005 from http://news.bbc.co.uk/1/hi/in_depth/africa/2004/sudan/default.stm

- **Identify Intellectual Property.** Identify intellectual property that an author or an organization has copyrighted, licensed, patented, trademarked, or otherwise taken to preserve their rights to the material. Some web pages list the points of contact and terms of use information at the bottom of the site's homepage. When uncertain, intelligence personnel should contact their supporting Judge Advocate General office before publishing information containing copyrighted or similarly protected intellectual property.

Open source research, coupled with an understanding of the COE, is the basis for an operational environment assessment. An operational environment assessment is a technique designed to apply the COE variables to a specific region, nation states, or non-state actors. It encompasses all the conditions, circumstances, and influences that affect the employment of military forces and the decisions of the unit commander. The operational environment assessment consists of a detailed examination and analysis of the eleven critical variables of the COE, their interaction and reciprocal relationships. Based on this analysis, the operational environment assessment identifies trends and issues with which units may have to grapple during their planning, preparation for, and execution of operations. As an unclassified document in whole or part), the operational environment assessment also serves as a useful tool for individual and collective training during preparation for operations in a specific area.

Every operational environment is complex, dynamic, and multi-dimensional.

An operational environment assessment provides a detailed look at a specific operational environment in terms of the eleven critical variables and their impacts. It identifies the critical relationships between the variables in the operational environment, how they affect one another, and how this affects military operations.

Some variables are dependent variables, whose value is determined by that of one or more other variables. For each dependent variable, the assessment identifies the most significant independent variables that are linked to it and shows their impact on the dependent variable under investigation.

To understand any operational environment, one needs to study and understand the synergy and interaction of variables and their reciprocal influence on one another. Within the analysis by variables, the operational environment assessment identifies key actors (nation-state and non-state) and assesses their impact on the operational environment.

This analysis of variables and actors helps to identify relevant trends and issues in the operational environment over time. Given the dynamic and fluid nature of the operational environment under investigation, an operational environment assessment requires continuous updates and additions in order to remain current and relevant.

2. CRITICAL VARIABLES

Open source research must address the eleven critical variables that describe the conditions in the potential operational environment. Collectively, these variables provide a complete framework for thoroughly assessing and understanding the complex and ever-changing combination of conditions, circumstances, and influences that affect military operations in any given real-world operational environment. While these variables can be useful in describing the overall (strategic) environment, they are most useful in defining the nature of specific operational environments. The variables do not exist in isolation from one another. The linkages of the variables cause the complex and often simultaneous dilemmas that a military force might face. Only by studying and understanding these variables—and their dynamic and complex combinations and interactions—will the US Army operational and tactical forces be able to keep adversaries from using them against them or to find ways to use them to its own advantage.

The eleven critical variables shown in Figure 1 are discussed below.

- **Physical Environment.** The physical environment defines the physical circumstances and conditions surrounding and influencing military forces and the execution of operations. The defining factors include urban settings and other complex terrain, all relevant infrastructures, weather, topography, hydrology, and environmental concerns.

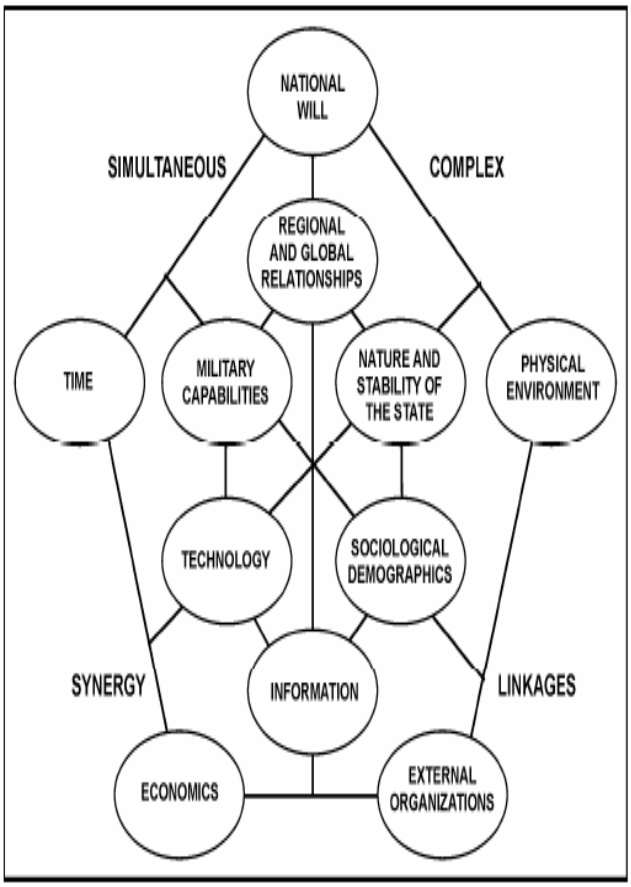

Figure 1: Critical variables of the operational environment

- **Nature and Stability of the State (or Other Critical Actors).** It is important to understand the nature and stability of the state or states with which or in which military operations take place. This variable, however, refers to the internal cohesiveness of the various political actors (nation states as well as non-state actors) with respect to the population, economic infrastructures, political processes, military and/or paramilitary forces, authority, goals, and agendas.
- **Sociological Demographics.** Sociological demographics refer to the traits and trends that have an impact on the human population of a particular group, area, country, or region. This includes its cultural, religious, and ethnic makeup.
- **Regional and Global Relationships.** Regional and global relationships include political, economic, military, or cultural mergers and partnerships. An actor's membership in or allegiance to such a relationship can determine its actions in terms of support, motivation, and alliance construct.

- **Military (or Paramilitary) Capabilities.** The military or paramilitary capabilities of various actors in the operational environment are a key concern. This variable includes such factors as equipment, manpower, training levels, resource constraints, and leadership issues. The military variable interacts with the other variables, and all the other variables can affect military and paramilitary capabilities.
- **Technology.** Technology represents the level or sophistication of technologies an actor could bring to the operational environment. Their level of integration and exploitation, and any niche technologies are important.
- **Information.** Information involves the access, use, manipulation, distribution, and reliance on information technology systems, both civilian and military, by a nation-state or non-state entity.

Information technology is the systems or mechanisms for preserving or transmitting information.

- **External Organizations.** External organizations refer to those entities found in an operational environment, which come from outside the confines of that specific operational environment but could impact the battlefield and related battle space. Such impact could be both positive and negative in nature—at the strategic, operational, and tactical levels and across the entire spectrum of conflict. An understanding of each group's varying and dynamic agendas, media philosophies, and international connections can be critical to the success of any military endeavor.
- **National Will (or Actors' Will).** Will encompasses a unification of values, morals, and effort between the population, the leadership or government, and the military or paramilitary forces. Through this unity, all parties are willing to sacrifice individually for the achievement of the unified goal. The interaction of military actions and political judgments, conditioned by national will, further defines and limits the achievable objectives of a conflict, thereby determining its duration and conditions of termination. It is imperative to study not just the national will of the state actors but also the will of the non-state actors (such as ethnic groups, political groups, insurgents, terrorist groups, and criminal organizations) involved in the operational environment. The will of non-state actors often affects the environment more significantly.
- **Time.** The time available for commanders to accomplish missions is determined by the goals and associated milestones established by the national political leadership. It is within this "timeframe" that all the elements of power—diplomatic, informational, military, and economic—must operate to achieve national objectives. How much time is available and how long events might take will affect every aspect of military planning, to include force package development, force flow rate, quality of intelligence preparation of the AO, and the need for forward-deployed forces and logistics. Time is often in favor of actors other than the US and its friends and multinational or coalition partners. Such actors often can afford to prolong the conflict and try to outlast the US will to continue operations in a particular operational environment.
- **Economics.** There may be significant differences among nation-states, organizations, or groups, regarding how they produce, distribute, and consume goods and services. Being able to affect another actor, positively or negatively, through economic rather than military means may become the key to regional hegemonic status or dominance. Economic deprivation is also a major cause of conflict. One actor may have economic superiority over another for many reasons, including access to natural resources or energy. Control of and access to natural or strategic resources can cause conflict. Military personnel operating in this complex environment may need to look beyond political rhetoric to discover a fundamental economic disparity among groups.

-6. Variables are fluctuating factors or elements that make up an operational environment. When operational, they define the conditions, circumstances, and influences that affect the employment of military forces and influence the options and decisions of the commander. The starting point for understanding the operational environment are those critical variables that reside in all operational environments and have the greatest impact on the military. See FM 7-100 for detailed information on the critical variables and operational environments.

An operational environment assessment provides a methodology for examining and understanding any potential operational environment. In effect, this assessment is an application of the COE concept to the specific operational environment under investigation. The methodology involves the following steps:

- **Define Variables.** Defining a variable simply means describing the nature and composition of each of the eleven variables in a specific operational environment.
- To help focus and facilitate the research effort, it is necessary for analysts to break down each variable into its subcategories of information (main topics and subtopics). This topic outline defines the scope and focus of the variable and serves as a guide for research on the variable in question.
- All eleven variables are present in all operational environments, but different operational environments typically will require different outlines within the variables. For example, a landlocked operational environment will not

require discussion of coastlines or ports under the Physical Environment variable or of naval forces under Military Capabilities. As analysts begin to populate the outline with information gleaned from their research, further refinements and additions to the headings and subheadings may be necessary.

- Because of the interrelated and sometimes overlapping nature of the individual variables, some subsets of available information may have a place under more than one variable. For instance, information technology might be addressed under both Information and Technology and could have an impact on several other variables.
- Analysts may determine links between the subcategories of one variable and the same, similar, or related subcategories that may exist under other variables. Thus, it may be necessary for one variable description to repeat some information contained under another variable or to cross-reference or provide an electronic link to it. Such linkages may be obvious when constructing the original topic outlines for the variables, or may become evident later—when analysts are populating the outlines with information.
- **Populate the Variable Outlines.** Analysts then conduct extensive research to populate the outline for each variable.
- This can be done with relevant information gleaned from all available open sources. The various sources can include official government documents, think-tank products, academic journals, open source periodicals, foreign press, websites, and interviews or discussions with various subject matter experts. Each variable is described as it applies to the specific operational environment in question. This step is an ongoing process, involving continuous updates as new or better information becomes available or when conditions change.
- Much of the most useful information about the potential operational environment may be available from open sources. An operational environment assessment can reveal key areas where information gaps exist. These gaps may become PIRs during the planning and the execution of operations—to be targeted by further open source research or perhaps by scarcer, more sensitive intelligence means. The operational environment assessment and underlying data form the basis of the GMI database and continuation of the operational environment assessment process when the unit deploys to the AO—possibly layered with additional information from classified sources.
- **Analyze Relationships, Linkages, and Trends**. The next step highlights the key cause-and effect relationships and linkages among the variables.
- In any operational environment assessment, the key to understanding the significance of the variables is to understand the relationships among the variables and how these affect military operations in the selected operational environment. Therefore, analysts can develop a matrix for each dependent variable that shows its critical relationships to other variables. This makes it possible to analyze each dependent variable from the perspective of its relationship or connectivity to other, independent variables. From this relational analysis, critical trends and issues become more evident.
- Clearly it is impossible to show every potential linkage and trend. Analysts should, however, identify and examine the most significant independent variables (linked to the dependent variable) to show their specific relationships to and impacts on the dependent variable under investigation. For instance, the variable of military capabilities is dependent on or influenced by virtually all the other variables.

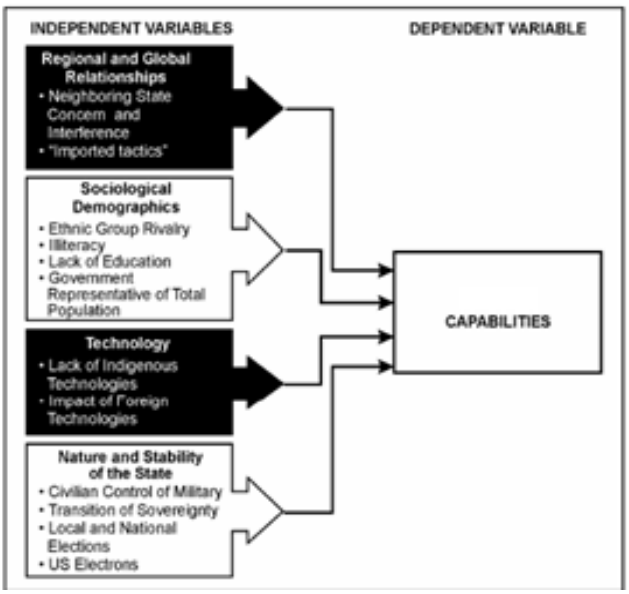

Figure 2: Example of Possible Links between Dependent and Independent Variables.

- Identify Key Facts and Impacts. Finally, analysts identify and highlight key facts and potential operational impacts for each variable.
- From the definition of variables and relationship analysis, analysts can attempt to identify trends over time. This trends analysis can provide an understanding of the dynamics of the variables and their impacts in a selected operational environment.
- Analysis can also identify possible trigger events in the operational environment based on relationships of variables across time.

3. EVALUATE INFORMATION

Deception and bias are of particular concern in OSINT operations. Secondary sources such as publications and

others who publish or broadcast information can intentionally or unintentionally add, delete, modify, or otherwise filter the information they make available to the public. It is important to evaluate the reliability of open sources in order to distinguish objective, factual information from that lacking merit, containing bias, or is part of an effort to deceive the listener or viewer.

Analysts evaluate each new item of information with respect to the reliability of the source and the credibility of the information (Table 4). An alphanumeric rating is assigned to each piece of information to indicate the degree of confidence the evaluator places on the information.

This rating is based on the subjective judgment of the evaluator and the accuracy of previous information produced by the same source.

The capabilities and performance of the collection resource may also be a factor in the evaluation. Intelligence personnel must assess the reliability of the source and the credibility of the information independently of each other to avoid the possibility of one factor evaluation biasing the other.

TYPE	DESCRIPTION	FACTORS
Primary Source	Has direct access to the information and conveys the information directly and completely.	**Access**. Did the source have direct access to the even or information? **Mediation**. Does the source provide a direct and complete view of the event or information? **Responsibility**. The more immediately, continuously, and directly a controller controls a source, the more responsible the source is to its controller and the more authoritative it is in presenting the controller's view. **Track Record**. Analysis of the source based on source's past behavior.
Secondary Source	Conveys information throught various types of filters: • Uses intermediary sources. • Summarizes, paraphrases, or excerpts. • Translates from the vernacular.	
Authoritative Source	Accurately reports information because it is known to be accountable to or has a track record demonstrating accuracy in reporting information from the leader, government, ruling party, or other element.	

Table 4: Types of Sources.

Source reliability ratings range from A (Reliable) to F (Cannot Be Judged) as shown in Table.4. If the source is new, they rate the source as F (Cannot Be Judged). An F rating does not necessarily mean the source is unreliable but that the collection and processing personnel have no previous experience with the source upon which to base a determination.

Table 5. Information credibility ratings range from 1 (Confirmed) to 8 (Cannot Be Judged) as shown in Table 5. If the information is new, they rate the content as 8 (Cannot Be Judged). An 8 rating does not necessarily mean the information is not credible but that the collection and processing personnel have no means of verifying the information.

CODE	RATING	DESCRIPTION
A	Reliable	No doubt of authenticity, trustworthiness, or competency; has a history of complete reliability; usually demonstrates adherence to known professional standards and verification processes.
B	Usually Reliable	Minor doubt about authenticity, trustworthiness, or competency; has a history of valid information most of the time; may not have a history of adherence to professionally accepted standards but generally identifies what is known about sources feeding any broadcast.
C	Fairly Reliable	Doubt of authenticity, trustworthiness, or competency but has provided valid information in the past.
D	Not Usually Reliable	Significant doubt about authenticity, trustworthiness, or competency but has provided valid information in the past.
E	Unreliable	Lacking in authenticity, trustworthiness, and competency; history of invalid information.
F	Cannot Be Judged	No basis exists for evaluating the reliability of the source; new information source.

Table 5: Information Credibility.

4. ANALYZE INFORMATION

During analysis, intelligence personnel use a variety of analysis techniques to discern facts, indicators, patterns, and trends in information and relationships between variables. The techniques apply inductive or deductive reasoning to understand the meaning of past events and predict future actions. Each technique is based on facts, observations, or assumptions about the operational environment. Intelligence personnel are mindful of injecting US or US military cultural bias into their analysis, particularly their assumptions. FM 33.4 (FM 34-3) and FM 2-01.3 (FM 34-130) provide more information about intelligence analysis techniques and procedures.

Structure of Information Extraction

CODE	RATING	DESCRIPTION
1	Confirmed	Confirmed by other independent sources; logical in itself; consistent with other information on the subject.
2	Probably True	Not confirmed; logical in itself; consistent with other information on the subject.
3	Possibly True	Not confirmed; reasonably logical in itself; agrees with some other information on the subject.
4	Doubtfully True	Not confirmed; possible but not logical; no other information on the subject
5	Improbable	Not confirmed; not logical in itself; contradicted by other information on the subject.
6	Misinformation	Unintentionally false; not logical in itself; contradicted by other information on the subject; confirmed by other independent sources.
7	Deception	Deliberately false; contradicted by other information on the subject; confirmed by other independent sources.
8	Cannot Be Judged	No basis exists for evaluating the validity of the information.

Table 6: Information Credibility Rating.

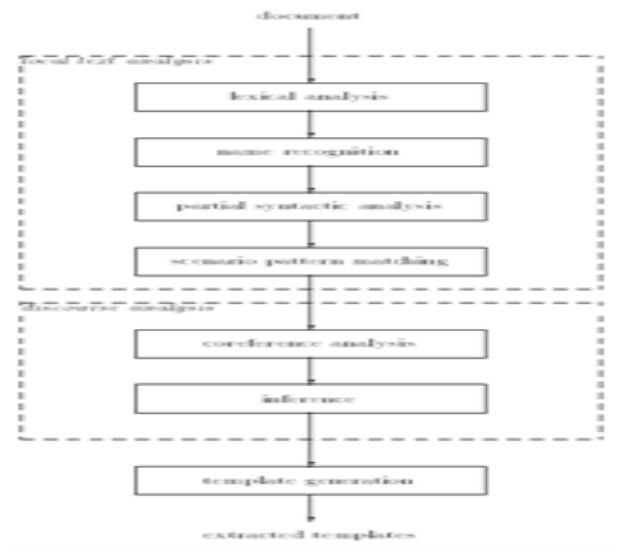

Figure 3: Information Extraction.

5. ASSOCIATION MATRIX

Intelligence analysts use the association matrix to establish known or suspected associations between individuals. Direct connections include, for example, face-to-face meetings or confirmed telephonic conversations. Figure 4 provides a one-dimensional view of the relationships and tends to focus on the immediate AO. Analysts can use association matrixes to identify those personalities and associations needing a more in-depth analysis in order to determine the degree of relationship, contacts, or knowledge between the individuals. The structure of the threat organization is formed as connections between personalities are made.

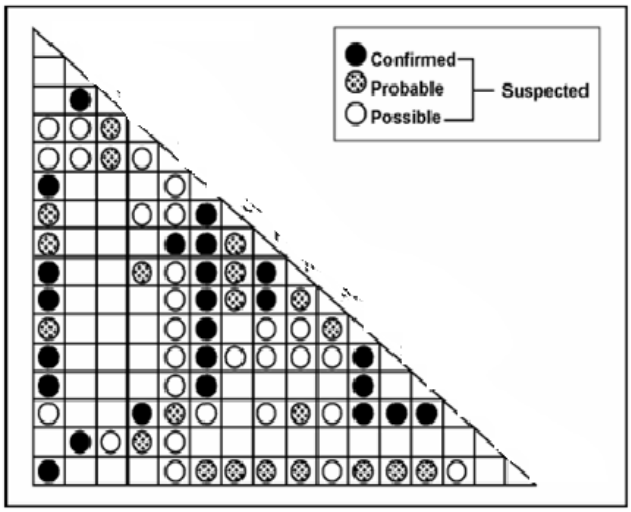

Figure 4: Example of an Association Matrix.

6. CONCLUSION

Investigators can now find answers to most of their elementary information needs using OSINT from professional software products Designed and built for that purpose. These support the complete intelligence lifecycle and user workflows based on established methodologies. They utilize core technology engines in the areas of information harvesting, data fusion, text analysis, link visualization, rules-based alerts and reporting in order to provide today's intelligence analyst a rich investigation environment and intuitive user experience. At the same time, these provide an excellent return on investment as users can generate critical intelligence with minimal effort.

REFERENCES

1. *Department of the Army Information Security Program*. 9 September 2000. Available online from Army Publishing Directorate at http://www.army.mil/usapa/epubs/index.html.

2. *Productions Requirements and Threat Intelligence Support to the US Army*. 28 June 2000. Available online from Army Publishing Directorate at http://www.army.mil/usapa/epubs/index.html.

3 Public Affairs Tactics, Techniques, and Procedures. 1 October 2000. Available online from Army Publishing Directorate at http://www.army.mil/usapa/doctrine/index.html

4. *Army Planning and Orders Production*. 20 January 2005. Available online from Army Publishing Directorate at http://www.army.mil/usapa/doctrine/index.html

5 *The Operations Process*. 31 March 2006. Available online from Army Publishing Directorate at http://www.army.mil/usapa/doctrine/index.html

6 *Opposing Force Doctrinal Framework and Strategy*. 1 May 2003.
FM 34-3. *Intelligence Analysis.* 15 March 1990.
FM 34-130. *Intelligence Preparation of the Battlefield*. 8 July 1994.
FM 34-3-61.1. *Public Affairs Tactics, Techniques, and Procedures.* 1 October 2000. Available online from Army Publishing Directorate.