Methods and Tools for Automated Data Collection and Collation of Open Source Information

Lisa Hagen
CAE Professional Services

Prepared for Dr. Bohdan L. Kaluzny
CJOC OR&A

DRDC CORA CR 2013–119
August 2013

Defence R&D Canada
Centre for Operational Research and Analysis

CJOC Operational Research and Analysis Team
Canadian Joint Operations Command

Methods and Tools for Automated Data Collection and Collation of Open Source Information

Lisa Hagen
CAE Professional Services

Prepared By:
CAE Professional Services
1135 Innovation Drive
Ottawa, Ont., K2K 3G7 Canada

Human Factors
Contractor's Document Number: No. 5453-001 Version 01
Contract Project Manager: Kevin Baker, 613-247-0342
PWGSC Contract Number: W7714-083663/001/SV
CSA: Bohdan L. Kaluzny, Team Leader, CJOC OR&A, 613-945-2392

Defence R&D Canada – CORA

Contract Report
DRDC CORA CR 2013-119
August 2013

Principal Author

Original signed by Lisa Hagen - CAE Professional Services

Lisa Hagen

Consultant, Human Systems Integration, CAE Professional Services

Reviewed by

Original signed by Ms Isabelle Julien

Ms. Isabelle Julien

DRDC CORA Section Head, Land and Operational Commands

Approved for release by

Original signed by Paul Comeau

Paul Comeau

DRDC CORA Chief Scientist

Defence R&D Canada – Centre for Operational Research and Analysis (CORA)

Abstract

A search of open source resources was completed in order to understand the landscape with respect to software tools that are readily available to support the automated collection and collation of open source information. The intent is to create an understanding of potential technological options for supporting open source intelligence (OSINT) activities within the Canadian Joint Operational Command (CJOC). To that end, this preliminary search revealed the following high-level findings:

- There are numerous existing tools with a select number available free of charge to support the collection and collation of OSINT.

- Individual tools typically provide functionality to support the several phases of the Intelligence cycle (i.e., collection, collation, and analysis).

- Individual tools are generally tailored to handle a specific class of OSINT (i.e., media, geospatial); however, certain tools possess functionality to handle multiple classes of OSINT material.

Résumé

Un examen des ressources ouvertes a été effectué dans le but de comprendre la situation en ce qui concerne les logiciels disponibles pour soutenir la collecte et le regroupement automatiques de l'information de sources ouvertes. L'objectif est d'assurer une compréhension des possibilités technologiques visant à soutenir les activités du renseignement de sources ouvertes (OSINT) au sein du Commandement des opérations interarmées du Canada (COIC). À cette fin, cet examen préliminaire en est arrivé aux conclusions de haut niveau suivantes :

- Il existe de nombreux outils, dont un certain nombre sont gratuits, permettant de collecter et de regrouper l'OSINT.

- Les outils offrent généralement des fonctions visant à soutenir les différentes phases du cycle du renseignement (collecte, regroupement et analyse).

- Les outils sont généralement adaptés au traitement d'une catégorie précise d'OSINT (médiatiques ou géospatiaux, par exemple), mais certains outils sont en mesure de traiter plusieurs classes d'OSINT.

This page intentionally left blank.

Executive summary

Methods and Tools for Automated Data Collection and Collation of Open Source Information

Hagen, L.; DRDC CORA CR 2013-119; Defence R&D Canada- CORA; August 2013.

This document is a deliverable for the project “Methods and Tools for Automated Data Collection and Collation of Open Source Information.” There were four tasks associated with this project and each is discussed below.

The first task was to perform a literature review to identify and list existing methods, tools, and software for automated data collection and collation of Open Source Intelligence (OSINT). This review was limited to publications from 2003 onwards and included searching Defence reports, Research and Development (RAND) reports, Defence Research and Development (DRDC) reports, journal articles, conference proceedings, websites, software reviews/manuals and other relevant compilations. The results of the literature review are presented in an annotated bibliography. The abstracts from each of relevant documents were used to summarize the pertinent details for each individual document. As part of this objective, selected papers (as chosen by the Technical Authority) were summarized or further investigated. A summary is provided below the abstract for each of these papers.

The second task was to identify software tools pertinent to automated collection and collation. The search included software to collate multiple Rich Site Summary (RSS) feeds, software to scan and monitor social networking sites (e.g., Twitter, Facebook, blogs, etc.), text parsing software, multilingual search tools, and Geographic Information Systems (GIS) software. Subsequently, the third task involved summarizing the tools and software that were identified during the search. This summary includes sources, links to websites, platforms supported, links to user and product manuals and underlying algorithms when this information was available. Finally, a categorization schema of the software tools was generated as part of the fourth project task. The schema involved classifying all of the software tools across two dimensions: intelligence cycle and open source information types. Based on these categories, the following high-level conclusions can be formed:

- There are numerous existing tools with a select number available free of charge to support the collection and collation of OSINT.
- Individual tools typically provide functionality to support several phases of the Intelligence cycle (i.e., collection, collation, and analysis).
- Individual tools are generally tailored to handle a specific class of OSINT (i.e., media, geospatial); however, certain tools are capable of handling multiple types of OSINT material.

Sommaire

Methods and Tools for Automated Data Collection and Collation of Open Source Information

Hagen, L.; DRDC CORA CR 2013-119; R&D pour la défence Canada- CARO; août 2013.

Ce document constitue un produit livrable pour le projet « Méthodes et outils de collecte automatique de données et de regroupement de l'information provenant de sources ouvertes. Quatre tâches sont liées à ce projet, et elles font toutes l'objet de discussions ci-dessous.

La première tâche était d'effectuer un examen des documents en vue de cerner les méthodes, les outils et les logiciels de collecte et de regroupement automatiques de renseignements de sources ouvertes (OSINT) actuellement disponibles et d'en dresser la liste. Cet examen s'est limité aux publications remontant, au plus, à 2003, et a comporté un examen de rapports de la Défense, de rapports de recherche et de développement (RAND), de rapports de recherche et de développement pour la Défense (RDDC), d'articles de journaux, de conférences, de sites Web, de critiques/de manuels de logiciels et d'autres compilations pertinentes. Les résultats de cet examen sont présentés dans une bibliographie commentée. Les résumés de chaque document pertinent ont été utilisés pour obtenir les détails importants. Dans le cadre de cet objectif, certains documents (sélectionnés par l'autorité technique) ont été résumés ou ont fait l'objet d'un examen plus approfondi. Un sommaire est présenté après le résumé pour chaque document.

Pour la seconde tâche, on devait trouver des logiciels touchant la collecte et le regroupement automatiques. Ces recherches comprenaient des logiciels utilisés pour le regroupement de multiples fils RSS (Rich Site Summary), pour le suivi des réseaux sociaux (Twitter, Facebook, les blogues, etc.) et pour le parsage de textes, ainsi que des outils de recherches multilingues et des logiciels utilisant le système d'information géographique (SIG). Par la suite, la troisième tâche consistait en la création d'un résumé des outils et des logiciels cernés pendant les recherches. Ce résumé comprenait les sources, les liens vers les sites Web, les plateformes soutenues, les liens vers les manuels de l'utilisateur et du produit et les algorithmes sous-jacents, quand cette information était disponible. Finalement, un schéma de classification des logiciels a été créé dans le cadre de la quatrième tâche. Ce schéma présentait une classification de tous les logiciels selon deux axes : le cycle du renseignement et l'information de sources ouvertes. En se fondant sur ces catégories, nous pouvons en arriver aux conclusions de haut niveau suivantes :

- Il existe de nombreux outils, dont un certain nombre sont gratuits, permettant de collecter et de regrouper l'OSINT.
- Les outils offrent généralement des fonctions visant à soutenir les différentes phases du cycle du renseignement (soit la collecte, le regroupement et l'analyse).
- Les outils sont généralement adaptés au traitement d'une catégorie précise d'OSINT (médiatiques ou géospatiaux, par exemple), mais certains outils sont en mesure de traiter plusieurs classes d'OSINT.

Table of contents

List of tables

1 Introduction

This document is a deliverable associated with the project entitled “Review of Methods and Tools for Automated Data Collection and Collation of Open Source Information”. The purpose of this project was to conduct a literature search and review and develop an annotated bibliography of existing methods and tools for gathering and collating information from open source information. This report was completed by CAE Inc. under Task #151 for contract W7714-083663/001/SV to Defence Research and Development Canada (DRDC) Centre for Operational Research and Analysis (CORA).

1.1 Background

The Canadian Joint Operations Command (CJOC) Operational Research and Analysis Team requested a literature search and review of software applications for automated data collection, collation and classification. The literature review incorporated both Canadian and foreign studies and tools related to the following areas: data collection and collation methods, tools and software.

1.2 Objective

The objective of this report was to conduct a literature search and develop an annotated bibliography of automated data collection, collation and classification methods and software tools.

1.3 This Document

This document is the deliverable required for this project. This document is structured accordingly:

- Section 1 – Introduction: Identifies the project background and objectives
- Section 2 – Method: Identifies the method used to conduct the literature search
- Section 3 – Annotated Bibliography and Data Collection Tools: Provides the annotated bibliography of the papers found during the literature search. The bibliographic reference and abstract are presented for each paper. The software tools are also described in this section
- Section 4 – Software Tool Classification - The open source intelligence software tools that were found during the literature search have been categorized in accordance with a classification schema
- Section 5 – Conclusions and Recommendations
- References – Identifies all papers, proceedings and book chapters that are referenced in the annotated bibliography

2 Method

The literature search was conducted using the following online archives and databases:

- Google Scholar;
- Google Book;
- North Atlantic Treaty Organisation (NATO) Research and Technology Organisation (RTO) website;
- Canadian Defence Information Database (CANDID) website;
- Research and Development (RAND) reports website;
- Defense Technical Information Center (DTIC); and
- DRDC website.

Google Scholar produced the most exhaustive list of papers. The papers identified during the literature search consist of internal technical reports, contractual reports, academic dissertations, conference proceedings and peer-reviewed journal articles. Sources used in the papers chosen for the annotated bibliography were reviewed to see if any of those papers were pertinent for this report. All papers are open source documents.

The following keywords and keyword combinations were used in this literature search:

- Open source intelligence (OSINT);
- Rich Site Summary (RSS) feeds;
- Social Media;
- Software/Online tools for collection of open source intelligence;
- Software/Online tools for collation of open source intelligence;
- Software/Online tools for collection of RSS feeds;
- Software/Online tools for collation of RSS feeds;
- Software/Online tools for collection and collation of RSS feeds;
- Software/Online tools for collection of social media;
- Software/Online tools for collation of social media;
- Software/Online tools for collection and collation of social media;
- Software/Online tools for collection of Geospatial Intelligence;
- Software/Online tools for collation of Geospatial Intelligence; and
- Software/Online tools for collection and collation of Geospatial Intelligence.

3 Annotated Bibliography and Data Collection Tools

This section presents the results of the literature and software tools search according to the following categories:

1. **Data Collection and Collation Methods:** Processes, procedures and methods used to collect and collate open source data.

2. **Data Collection and Collation Tools – Papers and Proceedings**: Tools (e.g., algorithms, platforms, metrics, models, architectures, etc.) used for collecting and collating open source data.

3. **OSINT Software Tools:** Software applications to perform the following:

   a. **Data Collection and Collation Tools**: Fully developed software applications that collect and collate open source data from RSS feeds, newsfeeds, telephone records, websites, public health sites, newspapers, chat rooms and blogs. Some of these applications are commercially available.

   b. **Social Media Search Tools**: Fully developed software and online applications that collect and collate social media. These applications are either free or commercially available for purchase.

   c. **Geospatial Intelligence Software Tools**: Fully developed software and online applications that collect and collate open source geospatial intelligence. These applications are either free or commercially available for purchase.

Results for the first two categories culminated in a series of papers. For each paper, the abstract has been captured to provide insight into its contents. Although some papers were applicable to multiple categories, each paper is recorded in a single category where it was deemed most applicable. For the third, fourth and fifth categories, results are a series of website references with an overview of each individual software application. Much of the information regarding the software tools category was extracted directly from the vendor websites. A complete list of references can be found in Section 6.

3.1 Data Collection and Collation Methods

Thibault, G., Gareau, L. M., & Le May, F. (2007). Intelligence collation in asymmetric conflict: A Canadian armed forces perspective. *Proceedings of the 10th Annual Conference on Information Fusion, Quebec City, Quebec*. doi: 10.1109/ICIF.2007.4408115

Intelligence in the asymmetric environment appears to lend itself to relying more heavily on information coming from human sources which may provide a wealth of opportunistic intelligence. This kind of information is increasingly available in very large quantities and various formats. The intelligence community has the challenge of ensuring that this collected information/knowledge, which is, by its very nature, mostly unstructured, can actually be "packaged" efficiently so that it can be readily and efficiently exploited for all of its intelligence value. After revisiting the intelligence process and paying particular attention to collation and analysis, the problem of collation is exposed and potential avenues of solutions presented.

Ulicny, B., Baclawski, K., & Magnus, A. (2007). New metrics for blog mining. In N. Glance, N. Nicolov, E. Adar, M. Hurst, & F. Salvettii (Eds.), *Proceedings of the 1st International Conference on Weblogs and Social Media,* Boulder, CO, USA. doi:10.1117/12.720454

Blogs represent an important new arena for knowledge discovery in open source intelligence gathering. Bloggers are a vast network of human (and sometimes non-human) information sources monitoring important local and global events, and other blogs, for items of interest upon which they comment. Increasingly, issues erupt from the blog world and into the real world. In order to monitor blogging about important events, we must develop models and metrics that represent blogs correctly. The structure of blogs requires new techniques for evaluating such metrics as the relevance, specificity, credibility and timeliness of blog entries. Techniques that have been developed for standard information retrieval purposes (e.g. Google's PageRank) are suboptimal when applied to blogs because of their high degree of exophoricity, quotation, brevity, and rapidity of update. In this paper, we offer new metrics related for blog entry relevance, specificity, timeliness and credibility that we are implementing in a blog search and analysis tool for international blogs. This tools utilizes new blog-specific metrics and techniques for extracting the necessary information from blog entries automatically, using some shallow natural language processing techniques supported by background knowledge captured in domain-specific ontologies.

Boury-Brisset, A. C., Frini, A., & Lebrun, R. (2011, June). *All-source Information Management and Integration for Improved Collective Intelligence Production*. Paper presented at the 16th Annual International Command and Control Research and Technology Symposium – Collective C2 in Multinational Civil-Military Operations, Quebec City.

Intelligence analysts face an ever increasing amount of heterogeneous information from various intelligence, surveillance, and reconnaissance (ISR) sources, including data from sensors as well as human and open sources, provided in disparate multimedia formats and distributed across different systems. Further research is required for enhanced management and integration of all-source intelligence information, and collective production of intelligence products. To this end, a novel approach for global Intelligence, Surveillance and Reconnaissance (ISR) information management and integration is needed. In the paper, we present the vision of a net-centric virtual centralization for the storage and management of all ISR information and intelligence products, making use of standards for data and services. This includes metadata standards for enhanced description of intelligence information for their subsequent retrieval or discovery. Also, semantic technologies and ontologies are envisioned to facilitate ISR integration through mediation of different representation schemes. Initiatives related to geospatial data management are presented. We then derive an architectural framework that incorporates the components and services required for our vision. Such an environment aims at providing users with services to store, access, discover, and integrate intelligence information in support of all source analysis and sense-making activities. By maximising the use of emerging standards and semantic technologies, the approach should facilitate ISR information integration, collaboration, and interoperability in multinational civil-military operations contexts.

Gibson, S. D. (2011). ***Open source intelligence (OSINT): A contemporary intelligence lifeline.*** **(Doctoral dissertation). Retrieved from Cranfield Defence and Security, Shrvenham database (1826/6524).**

Traditionally, intelligence has been distinguished from all other forms of information working by its secrecy. Secret intelligence is about the acquisition of information from entities that do not wish that information to be acquired and, ideally, never know that it has. However, the transformation in information and communication technology (ICT) over the last two decades challenges this conventionally held perception of intelligence in one critical aspect: that information can increasingly be acquired legally in the public domain-'open source intelligence'. The intelligence community has recognised this phenomenon by formally creating discrete open source exploitation systems within extant intelligence institutions. Indeed, the exploitation of open source of information is reckoned by many intelligence practitioners to constitute 80 percent or more of final intelligence product. Yet, the resources committed to, and status of, open source exploitation belies that figure. This research derives a model of the high order factors describing the operational contribution of open source exploitation to the broader intelligence function: context; utility; cross-check; communication; focus; surge; and analysis. Such a model is useful in three related ways: first, in determining appropriate tasking for the intelligence function as a whole; second, as a basis for optimum intelligence resource allocation; and third, as defining objectives for specifically open source policy and doctrine. Additionally, the research details core capabilities, resources, and political arguments necessary for successful open source exploitation. Significant drivers shape the contemporary context in which nation-state intelligence functions operate: globalisation; risk society; and changing societal expectation. The contemporary transformation in ICT percolates each of them. Understanding this context is crucial to the intelligence community. Implicitly, these drivers shape intelligence, and the relationship intelligence manages between knowledge and power within politics, in order to optimise decision-making. Because open source exploitation obtains from this context, it is better placed than closed to understand it. Thus, at a contextual level, this thesis further argues that the potential knowledge derived from open source exploitation not only has a unique contribution by comparison to closed, but that it can also usefully direct power towards determination of the appropriate objectives upon which any decisions should be made at all.

Ríos, S. A., & Muñoz, R. (2012). Dark Web portal overlapping community detection based on topic models. In ***Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*****,** ***2.*** **doi: 10.1145/2331791.2331793**

A hot research topic is the study and monitoring of on-line communities. Of course, homeland security institutions from many countries are using data mining techniques to perform this task, aiming to anticipate and avoid a possible menace to local peace. Tools such as social networks analysis and text mining have contributed to the understanding of these kinds of groups in order to develop counter-terrorism applications. A key application is the discovery of sub-communities of interests which main topic could be a possible homeland security threat. However, most algorithms detect disjoint communities, which means that every community member belongs to a single community. Thus, final conclusions can be omitting valuable information which leads to wrong results interpretations. In this paper, we propose a novel approach to combine traditional network analysis methods for overlapping community detection with topic-model based text mining techniques. Afterwards, we developed a sub-community detection algorithm that allows each member to belong to more than one sub-community.

Experiments were performed using an English language based forum available in the DarkWeb portal (IslamicAwakening).

3.2 Data Collection and Collation Tools – Papers and Proceedings

Reid, E., Qin, J., Chung, W., Xu, J., Zhou, Y., Schumaker, R., & Chen, H. (2004). Terrorism knowledge discovery project: A knowledge discovery approach to addressing the threats of terrorism. In H. Chen, R. Moore D. Zeng & J. Leavitt (Eds.), *Lecture Note in Computer Science: Vol. 3073. Intelligence and Security Informatics*, 125-145. doi:10.1007/978-3-540-25952-7_10

Ever since the 9-11 incident, the multidisciplinary field of terrorism has experienced tremendous growth. As the domain has benefited greatly from recent advances in information technologies, more complex and challenging new issues have emerged from numerous counter-terrorism-related research communities as well as governments of all levels. In this paper, we describe an advanced knowledge discovery approach to addressing terrorism threats. We experimented with our approach in a project called Terrorism Knowledge Discovery Project that consists of several custom-built knowledge portals. The main focus of this project is to provide advanced methodologies for analyzing terrorism research, terrorists, and the terrorized groups (victims). Once completed, the system can also become a major learning resource and tool that the general community can use to heighten their awareness and understanding of global terrorism phenomenon, to learn how best they can respond to terrorism and, eventually, to garner significant grass root support for the government's efforts to keep America safe.

Carroll, J. M. (2005). OSINT analysis using adaptive resonance theory for counterterrorism warnings. *Artificial Intelligence and Applications,* 756-760.

Open Source Intelligence is an extremely valuable source of data for intelligence analysts in identifying and analyzing potential terrorism warnings and indicators. A key problem is making sense of this large amount of data in time to prevent a catastrophic situation like what occurred on September 11, 2001. These events might been mitigated had U.S. intelligence agencies had better information technology tools for analyzing the situation according to the report of Congress's Joint Inquiry into the events leading up to the Sept. 11 attacks. Improvements to Information and Communications Technologies (ICTs) are necessary to provide the Homeland Security with the proper support for their missions. Artificial Neural Networks (ANNs) have been around for over half a century and their biologically-inspired capability allows functionality similar to the human brain via simulated neurons that can make near-human choices. ANNs are in their genesis with future applications include finance, marketing, medicine and security in data mining. Data mining enables a large amount of data to be sifted and provide avenues to learn or generalize information about that data using feature extraction. Adaptive Resonance Theory may provide another tool for this analysis.

Koltuksuz, A., & Tekir, S. (2006*).* Intelligence analysis modeling. In N. Szczuka, D. Howard, D. Slezak, H. Kim, T. Kim. I. Ko, G. Lee & P. Sloot (Eds.), *Advances in Hybrid Information Technology, 1,* 146-151. doi:10.1109/ICHIT.2006.253479

Intelligence is the process of supporting the policymakers in making their decisions by providing them with the specific information they need. Intelligence analysis is the effort of extracting the nature of intelligence issue with the policy goal in mind. It is performed by intelligence analysts who form judgments that add value to the collected material. With the increased open source collection capabilities, there has emerged a need for a model of intelligence analysis that covers the basic elements of valuable information: relevancy, accuracy, and timeliness. There exist models such as vector space model of information retrieval which only addresses the relevancy aspect of information and cannot cope with nonlinear document spaces. In this paper, we discuss the requirements of an integrated model of intelligence analysis along with its peculiar characteristics.

Neri, F., & Baldini, N. A. (2006, October). *Multilingual Text Mining based content gathering system for Open Source Intelligence*. Paper presented at the International Atomic Energy Agency (IAEA) Conference, Wien, Austria.

The number of documents available in electronic format has grown dramatically in recent years, whilst the information that governments provide to the IAEA is not always complete or clear. Independent information sources can balance the limited government-reported information, particularly if related to noncooperative targets. The process of accessing all these raw data, heterogeneous in terms of source and language, and transforming them into information is therefore strongly linked to automatic textual analysis and synthesis, which are greatly related to the ability to master the problems of multilinguality. This paper describes a multilingual indexing, searching and clustering system, designed to manage huge sets of data collected from different and geographically distributed information sources, which provides language independent search and dynamic classification features. The automatic linguistic indexing of documents is based on morpho-syntactic, functional and statistical criteria. The lexical analysis is aimed at identifying only the significant expressions in the whole raw text: the system parses each sentence, cycling through all possible sentence constructions. Using a series of word-relationship tests to determine the context, the system returns the proper meaning for each sentence. Once reduced to its part of speech (POS) and functional-tagged base form, later referred to its language independent entry of a subject-specific multilingual dictionary, each tagged lemma is used as a descriptor and becomes a candidate seed for clustering. The automatic classification of results is based on the Unsupervised Classification schema. By Multilingual Text Mining, analysts can get an overview of great volumes of textual data having a highly readable grid, which helps them discover meaningful similarities among documents and find any nuclear prolification and safeguard-related information. Providing automatic and language-independent features for document indexing and clustering, Multilingual Text Mining helps international agents cut through the information labyrinth and successfully overcome linguistic barriers.

Badia, A., Ravishankar, J., & Muezzinoglu, T. (2007). Text Extraction of Spatial and Temporal Information*. Proceedings of the Intelligence and Security Informatics Conference, USA,* 381. doi:10.1109/ISI.2007.379527.

Natural language analysis tools are very important for Intelligence tasks, since a considerable amount of information is available in documents of various types. The recent increase on use of OSINT has made documents even more abundant. Intelligence analysts require tools to help inspect, classify and analyze all this raw data. Situating documents (that is, finding their temporal and spatial coordinates) is vital to put events in the proper geo-strategical context; this in turn is an important part of the complex task of interpreting such events. Such information can help analysts relate events. In our project, we analyze documents at the sentence level. Each sentence is translated to a recursive structure anchored by a triple subject-action-object.

Summary: Dr. Badia was contacted via email and asked if he and his colleagues have further developed this program. He responded and indicated that this work was a former student's Master's thesis and the work has not been developed any further. He expressed an interest in collaborating to evolve the program if there was any interest in pursuing this objective.

Baldini, N. Neri, F., & Pettoni (2007*). A multilanguage platform for open source intelligence*. In A. Zanasi, C. Brebbia & N. Ebecken (Eds.), *Data Mining VIII: Data, Text and Web Mining and their Business Applications: Vol. 38* (pp. 18-20). New Forest, UK. doi:10.2495/DATA070321

Open Source Intelligence is an intelligence gathering discipline that involves collecting information from open sources and analyzing it to produce usable intelligence. The revolution in information technology is making open sources more accessible, ubiquitous, and valuable, making open intelligence at less cost than ever before. The explosion in OSINT is transforming the intelligence world with the emergence of open versions of the specialistic arts of human intelligence (HUMINT), overhead imagery intelligence (IMINT), and signals intelligence (SIGINT). The international Intelligence Communities have seen open sources grow increasingly easier and cheaper to acquire in recent years. But up to 80% of electronic data is textual and most valuable information is often hidden and encoded in pages which are neither structured, nor classified. The process of accessing all these raw data, heterogeneous in terms of source and language, and transforming them into information is therefore strongly linked to automatic textual analysis and synthesis, which are greatly related to the ability to master the problems of multilinguality. This paper describes a multilingual indexing, searching and clustering system, designed to manage huge sets of data collected from different and geographically distributed information sources, which provides language independent search and dynamic classification features. The Joint Intelligence and Electronic Warfare (EW) Training Centre (CIFIGE) is a military institute, which has adopted this system in order to train the military and civilian personnel of Defence in the OSINT discipline.

Summary: This tool appears to have been developed and implemented in the Joint Intelligence and EW Training Centre (CIFIGE) of the Italian Military. The tool, Stalker, is further discussed in this section (See Neri, F. & Pettoni. (2008) Stalker: A multilingual text mining search engine for open source intelligence). Based on a search of open source material, the tool does not seem to be commercially available.

Hopewell, P. H. (2007). *Assessing the acceptance and functional value of the Asymmetrical Software Kit (ASK) at the Tactical Level*. (Unpublished Master's Thesis). Naval Postgraduate School, Monterey, California.

The Asymmetrical Software Kit (ASK) is a software package built for U.S Army Special Operations Command (USASOC). It is designed to greatly expand and digitize the Intelligence Preparation of the Battlefield (IPB) process for Special Forces units. The purpose of this Thesis is to thoroughly evaluate the Tactical user's acceptance of this technological innovation. Technology Acceptance Model, which psychometrically measures users' perceptions of ease-of-use and utility to predict their intention to use the software, was applied in this analysis. The test population for this user acceptance survey is the Tactical (Group and below) level user of the ASK. These are the Special Forces Intel Sergeants (18Fs) on the Special Forces A-Teams (ODAs), and the Military Intelligence personnel at the Battalion and Group S2 (Staff Intelligence) sections. Respondents completed an anonymous, online survey on their impressions of the

ASK. The questions were focused on system usability and user acceptance in a military setting. Overall, the models used in this study showed an acceptable level of fit with the Tactical end-user's usability and acceptance assessments and exhibited satisfactory explanatory power. Users showed marked trends in response to questions concerning training, command involvement, and system availability. Qualitative input included a number of responses about the idiosyncrasies of certain programs, and the lack of high speed computers to run complex GIS queries. The findings from this study should provide some valuable insights to Program Managers about systems evaluation, and clarify how USASOC can design full spectrum software fielding to foster technology acceptance and use at the Tactical level.

Memon, N., Hicks, D. L., & Larsen, H. L. (2007). Harvesting Terrorists Information from Web. *Proceedings of the 11th International Conference on Information Visualization, Switzerland, IV07,* 664-671. doi: 10.1109/IV.2007.60

Data collection is difficult to do in any network analysis because it is hard to create a complete network. It is not easy to gain information on terrorist networks. It is a fact that terrorist organizations do not provide information on their members and the government rarely allows researchers to use their intelligence data. Very few researchers collected data from open sources, and to the best of our knowledge, no knowledge base is available in academia for the analysis of the terrorist events. To counter the information scarcity, we at Software Intelligence Security Research Center, Aalborg University Esbjerg Denmark, designed and developed a terrorism knowledge base by harvesting information from authenticated websites. In this paper we discuss data collection and analysis results of our ongoing research of Investigative Data Mining (IDM). In addition, we present a system architecture of our analyzing, visualizing and destabilizing terrorist networks prototype, i.e., iMiner, and also describe how we collected terrorist information using an Information Harvesting System.

Zanasi, A. (2007). *New forms of war, new forms of intelligence: Text mining.* Paper presented at the Information Technology for National Security Conference, Riyadh, Saudia Arabia.

This paper discusses text mining technologies, which allow the reduction of information overload and complexity, and analyzing texts in languages other than English.

Best, C. (2008). Web mining for open source intelligence*. Proceedings of the 12th International Conference on Information Visualization*, *England*, *IV08*, 321-325. doi:10.1109/IV.2008.86

Web mining for open source intelligence is the retrieval, extraction and analysis of information from on-line Internet sites. There are two separate applications areas this paper will review, namely live news-monitoring and targeted topic based data mining. Most newspapers and news agencies have Web sites with live updates on unfolding events, opinions and perspectives on world events. Most governments monitor news reports to feel the pulse of public opinion, and for early warning of emerging crises. The Joint Research Centre has developed significant experience in Internet content monitoring through its work on media monitoring (EMM) for the European Commission. EMM forms the core of the Commission's daily press monitoring service. Intelligence services and law enforcement agencies also require specific site monitoring and topic monitoring, and EMM technology has been applied to the wider Internet for this purpose. The software extracts and downloads all the textual content from monitored sites and applies information extraction techniques. These tools help analysts process large amounts of documents to derive structured data. Lastly the visualisation of the extracted data is important for analysts to identify patterns and trends derived from both news reports and Web mining.

Summary: This software has been further developed and is discussed in Section 3.3.1.1 - Europe Media Monitor Open Source Intelligence SUITE.

Kallurkar, S. (2008). *Targeted information dissemination.* (Report No. Unknown). Retrieved from the Defense Technical Information Center (DTIC) Website: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA480150.

Quantum Leap Innovations (QLI) developed a Targeted Information Dissemination (TID) system for rapid gathering and dissemination of the right information to the right people at the right time. The TID user interface shows tasks of an analyst. A hierarchical view of interests learned over a period of time is shown for each task. A table displays documents filtered-in by the user agent. The filtering is based on an interest profile that the agent manages on behalf of the user. The user can view and change the degree of filtering, document relevance and the interests related to task at any time. QLI focused their system to derive an early warning system (EWS) posed by a potential pandemic influenza episode, but the technology will be broadly applicable and configurable as an EWS for any future biological incident.

Neri, F. & Pettoni, M. (2008) *Stalker, a multilingual text mining search engine for open source intelligence. Proceedings of the 12th International Conference on Information Visualization*, *England, IV08,* 314-320. doi:10.1109/IV.2008.9

The revolution in information technology is making open sources more accessible, ubiquitous, and valuable. The international Intelligence Communities have seen open sources grow increasingly easier and cheaper to acquire in recent years. But up to 80% of electronic data is textual and most valuable information is often hidden and encoded in pages which are neither structured, nor classified. The process of accessing all these raw data, heterogeneous in terms of source and language, and transforming them into information is therefore strongly linked to automatic textual analysis and synthesis, which are greatly related to the ability to master the problems of multilinguality. This paper describes a content enabling system that provides deep semantic search and information access to large quantities of distributed multimedia data for both experts and general public. STALKER provides with a language independent search and dynamic classification features for a broad range of data collected from several sources in a number of culturally diverse languages.

Pfeiffer, M., Avila, M., Backfried, G., Pfannerer, N., & Riedler, J. (2008). Next Generation Data Fusion Open Source Intelligence (OSINT) System Based on MPEG7. *Proceedings of the Conference on Technologies for Homeland Security USA,* 41-46. doi:10.1109/THS.2008.4534420

We describe the Sail Labs Media Mining System which is capable of processing vast amounts of data typically gathered from open sources in unstructured form. The data are processed by a set of components and the output is produced in Moving Picture Experts Group (MPEG7) format. The origin and kind of input may be as diverse as a set of satellite receivers monitoring television stations or textual input from web-pages or RSS-feeds. A sequence of processing steps analyzing the audio, video and textual content of the input is carried out. The resulting output is made available for search and retrieval, analysis and visualization on a next generation Media Mining Server. Access to the system is web-based; the system can serve as a search platform across open, closed or secured networks. Data may also be extracted and exported and thus be made available in airgap networks. The Media Mining System can be used as a tool for situational awareness, information sharing and risk assessment.

Fei, Z., Xu, H., Weisheng, X., & Qidi, W. (2009). Analysis and Design of Web-Based Intelligence Mining Service System. *Proceedings of the Management and Service Science Conference, USA,* 1-4. doi:10.1109/ICMSS.2009.5300887

In this paper we have completed the analysis of intelligence mining system architecture framework and workflow, and showed the design frame of the web intelligence mining service system (IMSS). We have constructed and integrated intelligence experts brainpower supplemented by data mining technology, through the study on working mechanism of collection, analysis, services, counter-intelligence of the Web mining intelligence service system. We also expounded on the important and enlightening role that this research plays in the network security and the construction of military network.

Katakis, I., Tsoumakas, G., Banos, E, Bassiliades, N., & Vlahavas, I. (2009). An adaptive personalized news dissemination system. *Journal of Intelligent Information Systems: 32,* 191-212. doi: 10.1007/s10844-008-0053-8

With the explosive growth of the Word Wide Web, information overload became a crucial concern. In a data-rich information-poor environment like the Web, the discrimination of useful or desirable information out of tons of mostly worthless data became a tedious task. The role of Machine Learning in tackling this problem is thoroughly discussed in the literature, but few systems are available for public use. In this work, we bridge theory to practice, by implementing a web-based news reader enhanced with a specifically designed machine learning framework for dynamic content personalization. This way, we get the chance to examine applicability and implementation issues and discuss the effectiveness of machine learning methods for the classification of real-world text streams. The main features of our system named PersoNews are: (a) the aggregation of many different news sources that offer an RSS version of their content, (b) incremental filtering, offering dynamic personalization of the content not only per user but also per each feed a user is subscribed to, and (c) the ability for every user to watch a more abstracted topic of interest by filtering through a taxonomy of topics. PersoNews is freely available for public use on the WWW (http://news.csd.auth.gr). The current version of PersoNews is Beta and it currently allows users to monitor over 1920 feeds. These feeds cover a variety of areas such as blogs, conferences, science direct journals, and technology news. There is also a “various” category that contains feeds from news sources, universities, job sites, libraries, etc. Users can tailor the feeds they want to monitor and receive email reports regarding monitored feeds.

Neri. F, & Geraci, P. (2009). Mining textual data to boost information access in OSINT. *Proceedings of the 13th Conference on International Information Visualization, Spain, IV09,* 427-432. doi: 10.1109/IV.2009.99

The revolution in information technology is making open sources more accessible, ubiquitous, and valuable. The international Intelligence Communities have seen open sources grow increasingly easier and cheaper to acquire in recent years. Up to 80% of electronic data is textual and most valuable information is often encoded in pages which are neither structured, nor classified. The process of accessing all these raw data, heterogeneous for language used, and transforming them into information is therefore inextricably linked to the concepts of textual analysis and synthesis, hinging greatly on the ability to master the problems of multilinguality. This paper describes SYNTHEMA SPYWatch, a content enabling system for OSINT, which has been adopted by some Intelligence operative structures in Italy to support the Collection, Processing, Exploitation, Production, Dissemination and Evaluation cycle. By this system, operative officers can get an overview of great volumes of textual data, which helps them discover meaningful similarities among documents and find all related information.

Pouchard, L. C., Dobson, J. M., and Trien, J. P. (2009, March). ***A framework for the systematic collection of open source intelligence*****. Paper presented at the meeting of the Association for the Advancement of Artificial Intelligence Conference, Palo Alto, CA, USA**.

Following legislative directions, the Intelligence Community has been mandated to make greater use of Open Source Intelligence. Efforts are underway to increase the use of OSINT but there are many obstacles. One of these obstacles is the lack of tools helping to manage the volume of available data and ascertain its credibility. We propose a unique system for selecting, collecting and storing Open Source data from the Web and the Open Source Center. Some data management tasks are automated, document source is retained, and metadata containing geographical coordinates are added to the documents. Analysts are thus empowered to search, view, store, and analyze Web data within a single tool. We present ORCAT I and ORCAT II, two implementations of the system.

Summary: ORCAT I and II were developed by Oak Ridge National Laboratory located in Oak Ridge Tennessee. The software is open source software and can be downloaded free of charge from http://orcat.sourceforge.net/

Neri, F., Geraci, P., & Camillo, F. (2010*). Monitoring the Web Sentiment, The Italian Prime Minister's Case. Proceedings on the Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference, Denmark,* 432-434*.* doi:10.1109/ASONAM.2010.26

The world has fundamentally changed as the Internet has become a universal means of communication. The Web is a huge virtual space where to express individual opinions and influence any aspect of life. Internet contains a wealth of data that can be mined to detect valuable opinions, with implications even in the political arena. Nowadays the Web sources are more accessible and valuable than ever before, but most of the times the true valuable information is hidden in thousands of textual pages. Their transformation into information is therefore strongly linked to their automatic lexical analysis and semantic synthesis. This poster describes a Knowledge Mining study performed on over 1000 news articles or posts in forum/blogs, concerning the Italian Prime Minister Silvio Berlusconi, involved last year in the sexual scandal. All these textual contributions have been Morpho-Syntactically analysed, Semantically Role labelled and Clustered in order to find meaningful similarities, hilite possible hidden relationships and evaluate their sentiment polarity.

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E., & Zavarella, V. (2011). Online news event extraction for global crisis surveillance. In N. Nguyen (Ed.), *Lecture notes in Computer Science: Vol. 6910: Transactions on Computational Collective Intelligence,* 182-212. doi:10.1007/978-3-642-24016-4_10

This paper presents an endeavor aiming at construction of a real-time event extraction system for border security-related intelligence gathering from online news. First, the background and motivation behind the presented work is given. Next, the paper describes the event extraction processing chain, the specifics of the domain, i.e., illegal migration and related cross-border crime, and event moderation and visualisation aspects of the system.

Qureshi, P. A. R., Memon, N., Wiil, U. K., Karampelas, P., & Sancheze, J. I. N. (2011). Harvesting Information from Heterogeneous Sources. ***Proceedings from the European Conference on Intelligence and Security Informatics, Greece,*** **123-128. doi:10.1109.EISIC.2011.76**

The abundance of information regarding any topic makes the Internet a very good resource. Even though searching the Internet is very easy, what remains difficult is to automate the process of information extraction from the available online information due to the lack of structure and the diversity in the sharing methods. Most of the times, information is stored in different proprietary formats, complying with different standards and protocols which makes tasks like data mining and information harvesting very difficult. In this paper, an information harvesting tool (hetero Harvest) is presented with objectives to address these problems by filtering the useful information and then normalizing the information in a singular non hypertext format. We also discuss state of the art tools along with the shortcomings and present the results of an analysis carried out over different heterogeneous formats along with performance of our tool with respect to each format. Finally, the different potential applications of the proposed tool are discussed with special emphasis on open source intelligence.

Roberts, N. C. (2011). Tracking and disrupting dark networks: Challenges of data collection and analysis. ***Information Systems Frontiers, 13(1),*** **5-19. doi: 10.1007/s10796-010-9271-z**

The attack on September 11, 2001 set off numerous efforts to counter terrorism and insurgencies. Central to these efforts has been the drive to improve data collection and analysis. Section 1 summarizes some of the more notable improvements among U.S. government agencies as they strive to develop their capabilities. Although progress has been made, daunting challenges remain. Section 2 reviews the basic challenges to data collection and analysis focusing in some depth on the difficulties of data integration. Three general approaches to data integration are identified—discipline-centric, placed-centric and virtual. A summary of the major challenges in data integration confronting field operators in Iraq and Afghanistan illustrates the work that lies ahead. Section 3 shifts gears to focus on the future and introduces the discipline of Visual Analytics—an emerging field dedicated to improving data collection and analysis through the use of computer-mediated visualization techniques and tools. The purpose of Visual Analytics is to maximize human capability to perceive, understand, reason, make judgments and work collaboratively with multidimensional, conflicting, and dynamic data. The paper concludes with two excellent examples of analytic software platforms that have been developed for the intelligence community—Palantir and ORA. They signal the progress made in the field of Visual Analytics to date and illustrate the opportunities that await other information systems researchers interested in applying their knowledge and skills to the tracking and disrupting of dark networks.

Roy, J., & Auger, A. (2011, June). ***The multi-intelligence tools suite – Supporting research and development in information and knowledge exploitation.*** **Paper presented at the**

16th International Command and Control Research and Technology Symposium – Collective C2 in Multinational Civil-Military Operations, Quebec City, Canada.

While fulfilling its research mandate, the Intelligence and Information Section at DRDC Valcartier is constantly developing computer-based tools to support the analysts involved in intelligence activities. These tools are developed under different research projects, for various customers in diverse domains (e.g., improvised explosive devices and maritime situational awareness), to address specific aspects (e.g., the semantic analysis of unstructured documents, the use of automated reasoning to infer anomalous behaviours, etc.). For a large portion, they are built on knowledge-based systems technologies. However only providing stovepipe tools is not optimal; some integration is also required to create a synergy among them and facilitate the work of the analysts. The Multi-Intelligence Tools Suite (MITS) has thus been created as a federation of innovative composable and interoperable intelligence related tools, which are integrated and interleaved into an overall, continuous process flow relevant to the intelligence community. At the software system level, the backbone of the MITS is an integration platform built on open source Web services technologies, following services oriented architecture (SOA) design principles. The paper first reviews the main characteristics of the MITS. Then it discusses the central notions of domain knowledge and situational facts, describes the ingestion in the MITS of structured and unstructured data and information, briefly describes the main modules of the MITS, provides an exploitation example highlighting some of its powerful and innovative capabilities, and introduces the SOA platform and human-computer interaction components that constitute the MITS.

Noubours, S., & Hecking, M. (2012). Automatic exploitation of multilingual information for military intelligence purposes. *Proceedings of the Military Communications and Information Systems Conference (MCC), Poland*, 1-8.

Intelligence plays an important role in supporting military operations. In the course of military intelligence a vast amount of textual data in different languages needs to be analyzed. In addition to information provided by traditional military intelligence, nowadays the internet offers important resources of potential militarily relevant information. However, we are not able to manually handle this vast amount of data. The science of natural language processing (NLP) provides technology to efficiently handle this task, in particular by means of machine translation and text mining. In our research project Machine Translation for International Security Assistance Force (ISAF-MT) we created a statistical machine translation (SMT) system for Dari to German. In this paper we describe how NLP technologies and in particular SMT can be applied to different intelligence processes. We therefore argue that multilingual NLP technology can strongly support military operations.

Su, P., Li, D., & Su, K. (2012). An expected utility-based approach for mining action rules. *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics, China.* doi:10.1145/2331791.2331800

One of the central issues in data mining community is to make the mined patterns actionable. Action rules are those actionable patterns, which provide hints to a user what actions (i.e., changes within some values of flexible attributes) should be taken to reclassify some objects from an undesired decision class to a desired one. Both changing the value of a flexible attribute and the corresponding change of the value of a decision attribute may incur cost (negative utility) or bring benefit (positive utility) for the user. Obviously, the user is more interested in the rules which are expected to bring higher utility. In this paper, we formally define the expected utility of an action rule for measuring its interestingness. Our definitions explicitly state the problem of mining action rules as a search problem in a framework of support and expected utility. We also propose an effective algorithm for mining action rules with higher expected utilities. Our experiment shows the usefulness of the proposed approach.

Yang, H. C., & Lee, C. H. (2012, August). Mining open source text documents for intelligence gathering. In X. Jiang (Chair), *International Symposium on Technology in Medicine and Education*. Symposium conducted at the IEEE Sapporo Section, Hokkaido, Japan.

Intelligence collection and analysis always play a major role in a company's growth. Traditional intelligence management process was most concealed and required massive human effort. It also has the disadvantages of rarity and danger. Therefore OSINT emerged as a major intelligence collection and analysis approach. Differing from traditional approach, the sources of OSINT are publicly accessible and have the properties of openness and massiveness which may result in disadvantages of inconsistency and lack of validation. For now, most of the OSINT processing is conducted manually which requires massive human effort and time cost. Automatic processing of OSINT is then unavoidable for modern applications. Although there exists software services to aid such automatic processing, the functionality and degree of automation are still immature and limited. In this work we developed an automatic processing approach for OSINT based on proposed text mining techniques. This approach may automatically identify interesting events from various aspects from which business could benefit. The major contribution of this work is that we have developed high-order mining techniques for OSINT, which will benefit domains like national security, personal knowledge management, business intelligence, e-learning, etc.

3.3 OSINT Software Tools

In this section, software tools that are pertinent to automated data collection and collation of different types of open source information are presented. The technical attributes for each software tool have been summarized in a table at the beginning of each section and include:

1. **Application:** Name of the software;

2. **Uniform Resource Locator (URL) for Web-Enabled Applications:** The URL for accessing the web-enabled applications.

3. **Web-Based:** Applications that are created with Hyper Text Markup Language (HTML) and can be accessed with a browser. All of these applications are available free of charge.

4. **Windows-Based:** Applications that are operated within the Windows environment (as opposed to Linux, for example).

5. **Standalone:** Software that can work offline and does not necessarily require network connection to function. All of the software products that have been classified as standalone software must be purchased and downloaded from the vendor.

6. **Mobile:** Indicates if the web-enabled application or standalone software can be used with a mobile device.

7. **Free Software:** Specifies if the software can be downloaded for free. All the free software is Open Source Software.

8. **Free Trial Version:** Some software providers offer a free trial version of their software and this column shows which software products have a free trial period.

3.3.1 Data Collection and Collation Tools

In this section, software tools that collect and collate OSINT from RSS feeds, newsfeeds, telephone records, websites, public health sites, newspapers, chat rooms and blogs are summarized. Table 3-1 lists the tools that were investigated as part of this effort.

Table 3-1: Data Collection and Collation Tools

Application	Web-Enabled Applications or SW Downloads URL	Web Based	Windows Based	Stand-alone	Mobile	Free Software	Free Trial Version
EMM Newsbrief	http://emm.newsbrief.eu/NewsBrief/clusteredition/en/latest.html	Yes	N/A	N/A	No	N/A	N/A
EMM NewsExplorer	http://emm.newsexplorer.eu/NewsExplorer/home/en/latest.html	Yes	N/A	N/A	No	N/A	N/A
EMM MedSys	http://medisys.newsbrief.eu/medisys/homeedition/en/home.html	Yes	N/A	N/A	No	N/A	N/A
EMM Labs	http://emm-labs.jrc.it/	Yes	N/A	N/A	No	N/A	N/A
Palantir	http://www.palantir.com/solutions/intelligence/	No	Yes	Yes	No	No	No
Maltego	http://www.paterva.com/web6/sales/buy.php	No	Yes	Yes	No	No	No
Wynyard Group	https://wynyardgroup.com/downloads/solution/	No	Yes	Yes	No	No	No
Kapow	http://kapowsoftware.com/solutions/web-intelligence/osint.php	No	Yes	Yes	No	No	Yes
Rosette	http://www.basistech.com/government/osint/	No	Yes	Yes	No	No	Yes
Mario's Cyberspace Station	http://mprofaca.cro.net/	Yes	N/A	No	No	N/A	N/A
Webcase	http://veresoftware.com/i	No	Yes	Yes	No	No	Yes

Application	Web-Enabled Applications or SW Downloads URL	Web Based	Windows Based	Stand-alone	Mobile	Free Software	Free Trial Version
	ndex.php?page=order						

3.3.1.1 Europe Media Monitor Open Source Intelligence SUITE

Europe Media Monitor (EMM) Open Source Intelligence SUITE (http://ipsc.jrc.ec.europa.eu/events.php?idx=65) is a suite of software developed by the European Commission (DG JLS and Joint Research Center). These tools find, acquire, and analyse data from the Internet, including data which serves illegal activities such as the dissemination of child pornography, incitement to racist and xenophobic violence, provocation to commit terrorist attacks, recruitment into such groups, and training material to develop recruits into members. The tool can facilitate law enforcement agencies to cope with the volume and plurality of languages and information from the internet. The software can be downloaded free of charge by accessing the OSINT portal set up by the Commission services.

The European Commission's **Joint Research Centre** has developed a number of news aggregation and analysis systems to support European Union institutions and Member State organisations. The system was initially developed as an in-house application for the European Commission’s Directorate General Communication (DG COMM) to enhance their manual media monitoring and press cutting services. Since then, EMM has become a crucial instrument in the daily work of almost all Commission services and many other public organisations. EMM is the news gathering engine behind a number of applications. EMM monitors the live web, i.e. the part of the web that has ever changing content, such as news sites, discussion sites and publications. All applications are developed, maintained and run by the Joint Research Centre (information taken from http://emm.newsbrief.eu/overview.html#labs). The four publicly accessible **Web Portals developed by the European Commission are:**

- NewsBrief monitors and displays the hottest topics discussed during the past few minutes and hours with updates occurring every ten minutes. It monitors over 10,000 RSS feeds and HTML pages from 3750 news portals around the world plus 20 commercial newsfeeds. NewsBrief classifies all news according to hundreds of subjects and countries and users can be accessed on the web, via email and by RSS.

- NewsExplorer produces daily summaries of the news and also shows a map of where the day’s events occurred. It tracks stories over time and a calendar allows users to select past dates in order to view historic data. NewsExplorer links news stories across different languages allowing users to view different perspectives on the same story. Users can also access foreign language news regarding the same person, subject or event.

- MediSys displays only Public Health articles and groups them by disease or disease type. The tool monitors and analyzes the internet for threats that include communicable diseases, risks linked to chemical and nuclear accidents, and terrorist attacks. MediSys filters articles based on disease, symptoms, and chemical agents and these filters are based on users’ pre-defined word search combinations. Statistics for each category are provided and these statistics are also used to identify breaking news. Medical issues and trends are presented in graphs and users can be informed of news via email or Short Message Services (SMS).

- Labs provides users with access to the new advanced analysis systems resulting from information retrieval from the EMM data. From these data, themes and the number of articles from different countries are presented in different ways including World Map, Charts, and Tables. Lab monitors in real-time violent events and disasters and displays this information in both a browser and Google Earth. Users can also monitor social networks using Labs.

3.3.1.2 Palantir Intelligence

Palantir intelligence (http://www.palantir.com/solutions/intelligence) is a software solution employed throughout the American Intelligence Community to more efficiently, effectively, and securely exploit and analyze data, leading to more informed operational planning and strategic decision-making. Palantir Intelligence allows users to rapidly search through very large amounts of data and extract pertinent information. It then performs multi-dimensional data analysis to find hidden patterns, connections, trends, etc. Users can search and analyze information from enterprise data sources, network traffic, spreadsheets, telephone records, and email. Organizations can share data across data models and classification levels. Palantir Intelligence has version control that tracks all changes to the data. This allows multiple users to access the data without risking data integrity or security.

3.3.1.3 Maltego

Paterva (www.paterva.com) has developed a software tool, Maltego, for gathering open source intelligence. Maltego mines and gathers open source data and visually displays the results so that links and relationships are clearly defined. Maltego shows relationships between:

- People;
- Groups of people (social networks);
- Companies;
- Organizations;
- Web sites;
- Phrases;
- Affiliations;
- Documents and files; and
- Internet infrastructure.

Maltego also allows users to take bits of information and transform them into other entities. For example, a website address can be transformed into an Internet Protocol address and this will show up as a child to the original website entity (all information taken from the Patvera website). The following Maltego documentation is available on-line:

- User Guide: http://www.paterva.com/malv3/303/M3GuideGUI.pdf; and
- Maltego transforms user guides: http://www.paterva.com/malv3/303/M3GuideTransforms.pdf and http://www.paterva.com/malv3/303/Maltego3TDSTransformGuideAM.pdf.

3.3.1.4 Wynyard Group

Wynyard group (www.wynyardgroup.com) have a variety of software tools for gathering intelligence on financial and cyber-crime as well as tools for gathering open source and government intelligence. The OSINT tools collect, assess, understand and action publicly available intelligence. Wynyard OSINT and intelligence tools allow clients to collect open source intelligence, identify and extract entities, discover relationships, interrogate masses of structure and unstructured data for criminal intelligence and investigation purposes.

The Government Intelligence software is an integrated intelligence acquisition and discovery platform trusted by departments of homeland security, defense, justice and economic assistance. This tool provides advanced data collection, intelligence discovery and dissemination platform for homeland security, defense, justice and economic assistance, protecting against threat, crime and corruption (all information taken from the Wynyard website).

The highlights of the Wynyard government intelligence tool are:

- Flexible data collection and integration from multiple intelligence sources.
- Rapid data loading, updates and enrichment.
- Powerful text mining, link analysis, topic modelling, machine learning and prediction capabilities.
- Interactive, context-aware, highly personalised visualisation tools.
- Cross case, cross team and cross agency intelligence exchange or collaboration.
- Integration with Wynyard Financial Crime, Investigation Case Management and Digital Forensics solutions.

3.3.1.5 Kapow Software

Kapow software (www.kapowsoftware.com) is a web data extraction tool that allows users to harvest and integrate data OSINT data. After an OSINT capability has been developed, additional functionality can be provided to enhance capability as mission parameters change. Data harvesting, data integration and mission support capabilities are discussed below (extracted directly from the Kapow software white paper at http://kapowsoftware.com/assets/whitepapers/BuildingyourOSINTCapability.pdf):

- Data Harvesting is done using the Kapow Extraction Browser (used to support crawlers as they run) which uses a full JavaScript engine that allows the browser to see all the dynamic activity as a web page is being built. Kapow also handles cookies, session data, authentication data and other browsing artefacts allowing for a more complete extraction of data. Kapow allows the user to monitor and change the flow of data from a web page and this can be done in real time. Kapow Katalyst software tools allow users to conduct broad or surgical crawls. A broad crawl casts a wide net during the early stages of a search and if something of interest results from the broad search, a surgical crawl extracting additional contextual data from internal and external sources can be completed. Kapow Katalyst also has multiple language support for searching non-English web pages and captures and processes data from news sites, blogs, RSS feeds, and social media sites.

- Data Integration tools support the widest range of integration standards available. In addition to native support for Extensible Markup Language (XML), HTML, JavaScript Object Notation (JSON), Java Message Service (JMS), and Comma Separated Value (CSV), Kapow can also deliver results into Structured Query Language (SQL) databases and SML databases. Data can also be exposed as Web service Symbolic Optimal Assembly Program (SOAP) or Representation State Transfer (RESTful) and can callable via Java or .NET interfaces. In addition to invoking Web Services, Java or top level domain originally for network providers (.NET) applications during a crawl, it is possible to configure Kapow to directly invoke downstream enrichment capabilities during a crawl. This allows the user to decide based on mission parameters, and not technical limitations, whether real-time enhancement is necessary.

- Mission Support allows users to respond quickly to changes in the mission parameters and Kapow allows users the flexibility to respond quickly. For example, Kapow does not require a separate crawler for each site and sites that share similar characteristics (e.g., blogs) can use the same crawler because Kapow operates on the underlying Web page structure and not the presentation details of the site. Kapow also offers cloud functionality to ensure sufficient storage during periods of high uploads. Each crawler runs independently on its own thread and this provides linear scalability to support extremely high volume crawls.

3.3.1.6 Rosette

Basis technology (www.basistech.com) has a variety of commercial software solutions that extract intelligence from multilingual text. The software extracts intelligence from publicly available sources such as newspapers, blogs, chat rooms, Twitter, and other social media. The software, Rosette, can process volumes of data automatically, identify 55 languages, extract names of people and places and annotate names in foreign documents with English. The metadata from these searches can be used for applications such as link analysis, fact extraction, social media monitoring, and alerting systems.

3.3.1.7 Mario’s Cyberspace Station

Mario’s Cyberspace Station (http://mprofaca.cro.net/) is a web-enabled application that allows users to search the internet for global intelligence. The user can tailor the search by choosing one of four languages (English, Arabic, Hebrew or Croatian) and can search the entire web or employ filters such as Counterterrorism Blog, Middle East Intelligence Bulletin, Institute for Counterterrorism, and Global Security.

3.3.1.8 NewsNow

NewsNow (www.newsnow.co.uk) was developed in the United Kingdom and provides users with the ability to search thousands of media sites. There is a free online site and a subscription service. The online service allows users to enter keywords to search hundreds of news agencies located all over the world. Search results are lists of full text articles. The subscription service allows users to customize the software program and offers the following features (taken directly from http://www.newsnow.co.uk/services/):

- Tailored feeds: coverage options;

- 24/7 coverage of 42366+ sources in 20 languages from 146 countries;

- National, regional and international press;
- Consumer, trade and technical titles;
- TV news websites;
- Online magazines and newswires;
- Government and corporate announcements;
- Webzines, newsletters and blogs;
- Any additional sites added on request;
- Tailored feeds search options
  - Feeds built to your specifications, for your specific organisation's requirements;
  - Match articles only when given keywords occur within the same sentence, clause, paragraph or article;
  - Reject articles that come from the wrong sources, are in the wrong subject areas, or that specify irrelevant keywords or phrases;
  - Match 1, 10 or 100s of keywords and phrases simultaneously; and
  - Case-sensitive or case-insensitive, as required.
- Generic (pre-built) feeds
  - Choose from any of the 1,000s of pre-built topics already available on our website; and
  - Proven results: some of our website topics generate 100,000s page views per month.
- Delivery options
  - Within minutes of publication, on a fully-branded secure Client Portal;
  - Whether hourly or daily, by fully-branded 'email alerts' to one or more users; and
  - Fully-branded FTP and 'HTTP Push' to web or intranet servers.

3.3.1.9 Webcase

Veresoftware (http://veresoftware.com/) has developed a real-time forensic tracking software tool called WebCase that allows users forensically record IP addresses, chat sessions and other communication on the Internet. WebCase has the following features (taken directly from the website):

- Full screen and video capture documents every image, move and piece of high-jacked content;

- IP address identification and geolocation show where the website, blog or social network originates;

- Hash algorithms and date/time stamping, along with WebCase's evidence locker, preserve everything users need for an accurate case history;

- Manage numerous different targets and personas at once, so users can maintain clear case histories;

- XML file export into your most powerful analytic tools; and

- Easy generation and dissemination of collected data in a concise, consistent report directly to CD/DVD or tools of the Asymmetrical Software Kit (ASK).

3.3.2 Social Media Search Tools

Social Media Search Tools employ key words to conduct searches of social media (e.g., Twitter, Facebook, Tumblr, etc.) with results aggregated and often displayed in a graphical format. The following list of Social Media Search Tools was investigated.

Table 3-2: Social Media Search Tools

Application	Web-Enabled Applications or SW Downloads URL	Web Based	Windows Based	Stand-alone	Mobile	Free	Free Trial Version
Silobreaker	http://www.silobreaker.com/products/silobreaker-premium	No	Yes	Yes	No	No	Yes
Recorded Future	https://www.recordedfuture.com/2011/big-data-future-unlocking-predictive-power-web/	No	Yes	Yes	No	No	Yes
Social Mention	http://socialmention.com/	Yes	N/A	No	No	N/A	N/A
Addictomatic	http://addictomatic.com/	Yes	N/A	No	No	N/A	N/A
Whos Talkin	http://www.whostalkin.com/tools/	Yes	N/A	No	No	N/A	N/A
Kurrently	http://www.kurrently.com/	Yes	N/A	No	Yes	N/A	N/A
Samepoint	http://www.samepoint.com/	Yes	N/A	No	No	N/A	N/A
Newsnow	http://www.newsnow.co.uk/h/	Yes	N/A	No	Yes	N/A	N/A
Hootsuite	http://hootsuite.com/plan	No	Yes	Yes	Yes	No	Yes

Application	Web-Enabled Applications or SW Downloads URL	Web Based	Windows Based	Stand-alone	Mobile	Free	Free Trial Version
	s						
Trackur	http://www.trackur.com/options	No	Yes	Yes	No	No	Yes
Marketing Cloud	http://www.salesforcemarketingcloud.com/products/packages/	No	Yes	Yes	No	No	No
ORA and AutoMap	**AutoMap:** http://www.casos.cs.cmu.edu/projects/automap/downloads.php **ORA free software:** http://www.casos.cs.cmu.edu/projects/ora/software.php **ORA commercial license:** http://www.netanomics.com/products.html	No	Yes	Yes	No	Free Academic License and Commercial License to be purchased	Free Academic License and Commercial License to be purchased

3.3.2.1 Silobreaker

Silobreaker (http://www.silobreaker.com/) offers both a fully customizable and a downloadable suite of software for gathering and collating information from traditional and social media feeds. The customizable tool is called Silobreaker Enterprise Software Suite and the downloadable software intelligence gathering solution is called Silobreaker Premium (a free trial version of the premium software is available at the Silobreaker website). Each software package is discussed below and the information was taken directly from the Silobreaker website.

- Silobreaker Enterprise Software Suite has the following features:
  - Fully customizable in order to provide tailored solutions for language, translations, taxonomies, internal and external searches, etc.;
  - Add-on features such as real-time translation and voice transcription;
  - Capable of finding real-time and legacy data; and
  - Deals with the entire intelligence gathering workflow including back-end content aggregation, indexing, mining, classification and storage, analysis, user collaboration, report generation and decision support.
- Silobreaker Premium allows users to:
  - Monitor traditional and social media;

- Define and monitor targets;
- Trigger content aggregation;
- Extract connections between people, companies, events and places;
- Discover patterns in the news flow and track big movers;
- Perform Analyses;
- Display results in customized dashboards, email alerts or automatically generated reports; and
- Push results to colleagues via email alerts, RSS feeds or export findings into thirty party applications.

3.3.2.2 Recorded Future

Recorded future (https://www.recordedfuture.com/) is a downloadable software solution that continually scans thousands of media publications, blogs, government websites, and financial databases and then analyzes the text by identifying references to entities and events. These are organized by temporal expressions and, through online momentum and tone of language, are as having either positive or negative sentiment. Interactive tools allow users to analyze time patterns and have a better understanding of relationships and issues. A free 14-day trial can be downloaded from https://www.recordedfuture.com/2011/big-data-future-unlocking-predictive-power-web/

3.3.2.3 Social Mention

Social mention (http://www.socialmention.com/) is a social media search and analysis service that allows users to track and measure in real time topics on social media sites. Currently, Social Mention monitors over 100 social media sites including Twitter, Facebook, YouTube, Digg, Google, etc. The Social Mention Application Program Interface (API) enables developers to interact with the Social Mention web site programmatically. The API provides a single stream of real-time search data aggregated from numerous social media properties. It's designed to make it possible for anyone to access and integrate social media data into other applications. The API is free as long as fewer than 100 enquiries are made daily. If more than 100 enquiries are made on a daily basis, a monthly licensing fee is charged and this amount depends on the amount of API usage.

3.3.2.4 Addictomatic

Addictomatic (http://addictomatic.com/) is a free online service that searches the internet for news, blog posts, videos and images. Search results can then be customized using the dashboard feature. Searches are also customizable and cover topics such as top news, business and politics.

3.3.2.5 Whos Talkin

Whos talkin (www.whostalkin.com) is a social media search tool that allows users to search social media sites. The search and sorting algorithms combine data from more than 60 social media sites. In order to use this search tool, users download the free Whos Talkin iGoogle gadget and use this to conduct social network searches.

3.3.2.6 Kurrently

Kurrently (www.kurrently.com) is a free, real-time search engine that allows users to search Twitter and Facebook social networking sites. After a search term is entered, results are continually updated. Kurrently also offers a mobile app that can be used for searching Twitter and Facebook.

3.3.2.7 Samepoint

Samepoint (www.Samepoint.com) is a social media search tool that uses Syntax query language for tracking and filtering searches. Samepoint has the following features (taken from http://www.linkedin.com/company/samepoint-llc/samepoint-social-media-api-480983/product?trk=biz_product)

- One of the largest databases of social media web sites with 100-million+ sites and services, including YouTube, Wordpress, Tumblr, Blogspot, Facebook, Digg, millions of blogs and associated comments, bulletin boards, groups, videos, etc.;
- Ability to parse social media sites by type (i.e., blogs, social networks, bulletin boards, etc.) and provide analysis and visualization of breakdown of mentions;
- Real-time indexing for up-to-the-minute results;
- Advanced semantic analysis that automatically groups discussions by theme;
- Ability to "learn" where hot spots of conversations are taking place;
- Identify key metrics and establish benchmarks;
- Identify influencers;
- Identify influential sources where high-level of discussions are taking place; and
- On-the-fly sentiment analysis.

3.3.2.8 Hootsuite

Hootsuite (http://hootsuite.com/) has software tools that are both free and available for purchase. The tools allow users to search and manage RSS feeds and multiple social networks (e.g., Facebook, Twitter, Foursquare, Myspace, Mixi, etc.). Hootsuite allows users to monitor Facebook likes, comments and page activity from the dashboard and to conduct historical comparisons to see trends over time. Hootsuite also has analytics tools and customizable reports to enable users to view results in a variety of ways (e.g., graphs, charts, plots, etc.). Analytics tools include:

- The Google Analytics tool and URL parameters can track conversation and drill down into site traffic to allow users to see the source of conversation and the region where the conversation originated.

- The Twitter Analytics tool tracks the number of followers, following, lists, and mentions.

3.3.2.9 Trackur

Trackur (www.trackur.com) is software tool that allows users to monitor news sites, blogs, RSS feeds, social media (e.g., Twitter, Facebook, etc.), forums, images and videos and custom feeds. Results from searches can exported into excel or can be displayed with the Trackur RSS/XML feeds, through email alerts or on the users' dashboard. All social media conversations are archived which allows for a follow-up deeper analysis. Also, Trackur allows users to see names of people who are involved in social media conversations. Trackur can be used on a computer, tablet or mobile device.

3.3.2.10 Marketing Cloud

Marketing Cloud (http://www.salesforcemarketingcloud.com/) is a software tool that allows users to listen to conversations on the social web in real time. Additionally, Marketing Cloud also archives posts and they have a data base that contains more than 55 billion posts and dates back to May 2008. Marketing Cloud allows users to listen to conversations all over the globe in 19 different languages. With Marketing Cloud, users can dig deep into online conversations to learn about authors' demographics, whether a conversation is positive, negative or neutral and the emotion behind posts. Marketing Cloud uses text analytics and semantic technology to analyze the relationships between words in social posts which revels subtleties in social media users' intentions. Marketing Cloud lets users see in real-time what issues, people, places and things are being discussed and see who the online influencers are.

3.3.2.11 ORA and AutoMap

ORA and AutoMap are tools developed by the Center for Computational Analysis of Social and Organizational Systems (CASOS) at Carnegie Mellon University.

ORA (http://www.casos.cs.cmu.edu/projects/ora/) is a dynamic meta-network assessment and analysis tool developed by. It contains hundreds of social network, dynamic network metrics, trail metrics, procedures for grouping nodes, identifying local patterns, comparing and contrasting networks, groups, and individuals from a dynamic meta-network perspective. ORA has been used to examine how networks change through space and time, contains procedures for moving back and forth between trail data (e.g. who was where when) and network data (who is connected to whom, who is connected to where), and has a variety of geo-spatial network metrics and change detection techniques. ORA can handle multi-mode, multi-plex, multi-level networks. It can identify key players, groups and vulnerabilities, model network changes over time, and perform Course of Action (COA) analysis. It has been tested with large networks (10^6 nodes per 5 entity classes). Distance based, algorithmic, and statistical procedures for comparing and contrasting networks are part of this toolkit. Just as critical path algorithms can be used to locate those tasks that are critical from a project management perspective, the ORA algorithms can find those people, types of skills or knowledge and tasks that are critical from a performance and information security perspective. ORA can be applied both within a traditional organization and on covert networks. An academic license can be downloaded at http://www.casos.cs.cmu.edu/projects/ora/software.php or a commercial license can be purchased through Netanomics at http://www.netanomics.com/.

AutoMap (http://www.casos.cs.cmu.edu/projects/automap/) is a text mining tool that works seamlessly with ORA and allows for the extraction of network data from unstructured text. AutoMap can extract the following types of information:

- Content (concepts, words and frequencies);
- Semantic Networks (concepts and relationships);
- Meta-Networks (named entities and links); and
- Sentiment (attitudes and beliefs).

AutoMap contains both pre- and post-processors that “clean up” the text data. The pre-processor cleans the raw text data so that it can be processed by the post-processor tools. The pre-processors include tools such as a pdf to txt convertor and non-printing character removal. Pre-processing also reduces data into concepts and then statement formation rules determine how to link extracted concepts into networks. The post-processing tools include procedures that link to gazetteers (geographical dictionary or directory that contains important information about places and place names) and supplement the code with latitude and longitude information. There are also tools that create, maintain and edit delete lists.

3.3.3 Geospatial Intelligence Software Tools

Geospatial Intelligence software tools allow users to view, understand, manipulate and understand geographically referenced information. A variety of software tools are currently commercially available. Table 3-3 lists the tools that were investigated as part of this effort.

Table 3-3: Geospatial Intelligence Software Tools

Application	Web-Enabled Applications or SW Downloads URL	Web Based	Windows Based	Stand-alone	Mobile	Free	Free Trial Version
Esri ArcGIS Tools	http://www.esri.com/products	No	Yes	Yes	Yes	Yes - Open Source Software	Yes
PCI Geomatics	http://www.pcigeomatics.com/products/geomatica-2013sensible-packaging	No	Yes	Yes	No	No	Yes
Geographic Resources Analysis Support System (GRASS)	http://grass.osgeo.org/download/	No	Yes	Yes	No	Yes - Open Source e	N/A
Geospatial Intergraph	http://www.intergraph.com/sgi/industries.aspx	No	Yes	Yes	No	No	No
Jagwire	http://www.exelisinc.com	No	Yes	Yes	Yes	No	No

Application	Web-Enabled Applications or SW Downloads URL	Web Based	Windows Based	Stand-alone	Mobile	Free	Free Trial Version
	/solutions/Jagwire/Pages/default.aspx						

3.3.3.1 Esri ArcGIS

Esri ArcGIS (www.esri.com) is a platform that helps users to create solutions through the use of geographic information. They offer a variety of products to help with designing and managing solutions, data visualization and geographic intelligence and providing users with high-quality ready to use maps. These tools include the following:

- ArcGIS is a platform for designing and managing solutions through the application of geographic knowledge. ArcGIS allows users to create web applications for stakeholders and also allows users to access and utilize information that others have shared. The desktop application uses predefined maps that allow users to build models so that data can be analyzed scientifically. Users can also create geodatabases and manage and edit geographic data. There is also a mobile application for ArcGIS and this allows users to access the most up-to-date information. Field staff is able to use the mobile application to capture, update and analyze geographic information and share it with colleagues. ArcGIS allows users to upload and store maps and data in the cloud and this allows end users to access information without having to install software or worry about data management.

- Esri Location Analytics provides data visualization and geographic intelligence for business analytics and is not very applicable to military intelligence gathering.

- Esri Data provides users with the most current and accurate geospatial data available. Users can access imagery, street, shaded relief and topographic data as well as demographics, consumer spending and marketplace data.

Other companies have used ArcGIS tools to develop Geospatial Intelligence (GeoInt) tools for military and defence applications. Three of these tools are (taken directly from the Esri website):

- Distributed Geospatial Intelligence Network (DGInet): This technology provides an enterprise solution for geospatial intelligence data. It is a web-based enterprise geographic information system (GIS) and has been designed for use by novices and experts alike. The DGInet uses low bandwidth to allow users to search enormous amounts geospatial and intelligence data for data discovery, dissemination and horizontal fusion of data. The features DGInet provides are:

    - A scalable Java Web service environment within which Web services can be easily utilized, added, exposed, maintained, and integrated with collaborative geospatial capabilities;

    - A powerful architecture that will satisfy every agency and organization’s operational need for a geospatial enterprise system for dissemination within a robust collaborative environment;

    - A SOA accessible via portal-based browsers, applets, and heavy clients;

  - A collection of distributed Web services implemented as Java Web services;
  - Web map services and geoprocessing Web services across multiple organizations/nodes;
  - XML-based metadata broadcast search;
  - Selective data display/data fusion;
  - Data download capability; and
  - Data management services.

- GeoRover Tools for ArcGIS is a set of tools that streamlines the process of creating, importing, editing and sharing GIS data while easily allowing the user to perform real-time or post-collection processing of data from Global Positioning System (GPS) receivers, digital cameras, voice recorders and more. Additionally, this suite of tools imports text files, spreadsheets and data bases and can create interactive Web pages, spreadsheets and slide shows of the GIS and collected data.

- Communication System Planning Tools are tools that have been developed for the United States Department of Defense and can be used as a stand-alone tool or it can be integrated into and run within existing applications that are used for communication system analysis. To conduct this analysis, the tool incorporates existing and planned electromagnetic wave prediction models. The tool has three extensions which allows analysts to incorporate into maps frequencies between 150 kilohertz (KHz) to 2 megahertz (MHz), 2 MHz to 20 MHz, and from 20 MHz to gigahertz (GHz). This frequency data are plotted onto a map and the analyst can use this information to conduct analyses such as interference studies, overlap evaluations, point-to-point link analysis and coupled indoor/outdoor analyses.

3.3.3.2 PCI Geomatics

PCI Geomatics (http://www.pcigeomatics.com/) has developed a remote sensing software tool called Geomatica that allows users to load satellite and aerial imagery so that advanced analyses on this imagery can be performed. Geomatica 13 is the most current version of the software. The Geomatica software tool has a number of programs including (Information taken directly from the Geomatica Tutorial (http://www.pcigeomatics.com/pdf/tutorials/QuickStart_v10.pdf) and the Geomatica 13 flyer (http://www.pcigeomatics.com/pdf/Geomatica2013/Geomatica_2013_Flyer.pdf)):

- Focus is the main data visualization environment. It includes geospatial processing algorithms, data capture functionality, and information extraction and analysis tools;

- OrthoEngine is a complete photogrammetry environment offering geometric correction, orthorectification, Mosaicking, as well as 3-D visualization and data extraction environments;

- Modeler is a visual scripting environment that provides an interactive methodology for developing, automating and batch-processing both simple and complex workflows;

- EASI is a command-line-based scripting environment that provides workflow development as well as customization functionality;

- FLY! Is a visualization tool that renders perspective scenes and interactive 'fly-throughs' by using imagery and elevation information;

- Chip Manager lets users create and manage image chip libraries that are used as ground controls in orthorectification workflows;

- GeoRaster Metadata Mapper lets users store, index and manage geospatial files in Oracle 10g databases; and

- Atmospheric Correction allows users to automatically detect cloud and haze which makes it easier to create seamless mosaics in cloud areas.

3.3.3.3 Geographic Resources Analysis Support System (GRASS)

Geographic Resources Analysis Support System (GRASS) (http://grass.osgeo.org/) is an open source GIS software (GNU General Public License) that can be downloaded free of charge. The software was developed by the United States Army Construction Engineering Research Laboratories which is a branch of the United States Army Corp of Engineers. The software can be used for geospatial data management and analysis, processing images, producing graphics and maps, spatial modelling and visualization. The GRASS GIS capabilities are described below (taken directly from the website (http://grass.osgeo.org/documentation/general-overview/):

- **Raster analysis** provides automatic rasterline and area to vector conversion, Buffering of line structures, Cell and profile data query, Color table modifications, Conversion to vector and point data format, Correlation / covariance analysis, Expert system analysis , Map algebra (map calculator), Interpolation for missing values, Neighbourhood matrix analysis, Raster overlay with or without weight, Reclassification of cell labels, Resampling (resolution), Rescaling of cell values, Statistical cell analysis, Surface generation from vector lines;

- **3D-Raster (voxel) analysis** provides 3D data import and export, 3D masks, 3D map algebra, 3D interpolation (Regularised Splines with Tension), 3D Visualization (isosurfaces), Interface to Paraview and POVray visualization tools;

- **Vector analysis** provides contour generation from raster surfaces (Inverse Distance Weighting, Splines algorithm), Conversion to raster and point data format, Digitizing (scanned raster image) with mouse, Reclassification of vector labels, Superpositioning of vector layers;

- **Point data analysis provides** Delaunay triangulation, Surface interpolation from spot heights, Thiessen polygons, Topographic analysis (curvature, slope, aspect), and Light Detection and Ranging (LiDAR);

- **Image processing provides c**anonical component analysis (CCA), Color composite generation, Edge detection, Frequency filtering (Fourier, convolution matrices), Fourier and inverse fourier transformation, Histogram stretching, IHS transformation to Red, Green, Blue (RGB), Image rectification (affine and polynomial transformations on raster and vector targets), Ortho photo rectification, Principal component analysis (PCA), Radiometric corrections (Fourier), Resampling, Resolution enhancement (with RGB/IHS), RGB to

Intensity Hue and Saturation (IHS) transformation, Texture oriented classification (sequential maximum a posteriori classification), Shape detection, Supervised classification (training areas, maximum likelihood classification), Unsupervised classification (minimum distance clustering, maximum likelihood classification);

- **DTM-Analysis provides** contour generation, Cost/path analysis, Slop/aspect analysis, Surface generation from spot heights or contours;
- **Geocoding** provides geocoding of raster and vector maps including LiDAR point clouds;
- **Visualization** provides 3D surfaces with 3D query (NVIZ), Color assignments, Histogram presentation, Map overlay, Point data maps, Raster maps, Vector maps, Zoom / unzoom function;
- **Map creation provides** image maps, Postscript maps, HTML maps;
- **SQL-support** provides database interfaces (Dbase File [DBF], Structured Query Language Lite [SQLite], Postgre Structured Query Language [PostgreSQL], MY Structured Query Language [MySQL], Open Data Base Connectivity [ODBC]);
- **Geostatistics provides** interface to "R" (a statistical analysis environment), Matlab; and
- **Furthermore** provides erosion modelling, Landscape structure analysis, Solution transport, and Watershed analysis.

3.3.3.4 Geospatial Intergraph

Geospatial Intergraph (www.geospatial.intergraph.com) has GIS, remote sensing, photogrammetry tools as well as server-based products:

- GIS products allow users to utilize and manage data that is in the users' geographic information systems. With the GIS software products (GeoMedia, GeoMedia Smart Client and Geospatial Spatial Data Infrastructure), users can query multiple GIS and combine different geospatial resources into a single map view to obtain a clear understanding of the real-world scenarios they represent. The links between the Intergraph remote sensing and GIS products enable GIS updates from the users' raster imagery which improves the accuracy and relevance of the users' data resources.

- Remote Sensing products allow users to process data to enhance visibility of certain image elements and analyze it to extract information that cannot be detected from visual inspection along. These products help users:

    - Orthorectify (process an aerial photograph to geometrically correct it so that the scale of the photo is uniform) raster imagery;
    - Detect changes between two or more raster images;
    - Analyze imagery using spectral signatures;
    - Process and analyze radar data;
    - Classify imagery;

    - Create presentation products (Microsoft Word and PowerPoint documents) with imagery; and
    - Convert files.
- Photogrammetry products allow users to connect imagery to locations on the earth's surface and create accurate representation of the earth from remotely sensed data.
- Server products provide Service Oriented Architecture based enterprise solutions for managing and delivering geospatial data, services and workflows. These tools allow for:
    - Data Management;
    - Creating, publishing and analyzing web-enabled maps;
    - Manage and serve secure or licensed information using standards-based web services;
    - A portal that can be used for finding, viewing, querying, analyzing and consuming geospatial data and web services published by Intergraph or other third party products.

3.3.3.5 Jagwire

Jagwire (http://www.exelisinc.com/solutions/Jagwire/Pages/default.aspx) is a web-enabled product that allows users to very quickly access mission-critical geospatial intelligence data. Jagwire is built around a SOA that allows users to quickly deploy Jagwire into existing systems. Jagwire offers four different suites:

- Jagwire Enterprise Suite is designed for large scale Data Center environments, scaling to support large numbers of users, sensors and platforms along with long-term data archive applications.
- Jagwire Ground Suite is built for forward deployed locations and Ground Control Stations, providing assistance for specific tactical objectives. Scaled for reduced number of platforms and sensors.
- Jagwire Air Suite supports in-flight processing, storage and dissemination operations.
- Jagwire Mobile provides real time and archive access to critical intelligence data for deployed users.

4 Software tool classification

The OSINT software tools that were found during the literature search have been classified along two axes: steps of interest within the intelligence cycle and type of open source information.

4.1 Intelligence Cycle

Typically, the intelligence cycle is decomposed into the following four steps (Canadian Forces Joint Publication (CFJP) 2-0 Intelligence, 2011): direction, collection, processing, and dissemination. While there is a general progression through these four steps, these are not discrete events following a sequential process whereby one step is performed to completion prior to the next commencing. As such, these steps will occur in parallel.

For the purposes of this work, the Collection step as well as the Collation and Analysis sub-steps within the Processing step was used as part of the classification schema. These steps and sub-steps are defined as follows:

1. **Collection.** As the second step in the intelligence cycle, collection is “the exploitation of sources by collection agencies and the delivery of the information obtained to the appropriate processing unit for use in the production of intelligence” (CFJP, 2011, p. 3-5). It involves the collection of information and intelligence to meet the commander’s information. Generally, intelligence collection will be provided via two avenues: agencies (e.g., Joint Operation and Intelligence Center (JOIC) or semi-permanent establishments in foreign countries such as a deployed All Source Intelligence Center (ASIC)) and sources (e.g., human intelligence (HUMINT), SIGINT, OSINT)

2. **Processing.** Processing is the action of turning raw data into intelligence. Specifically, this step is defined as “sorting collected information and converting it into a form suitable for the production of intelligence” (CFJP, 2011, p. 3-6). Processing is a structured series of actions and can be further decomposed into the following sub-steps (CFJP, 2011):

   a. **Collation** is the step that consists of “procedures for receiving, recording, and grouping all information collected” (CFJP, 2011, p. 3-6). In practice, this involves the procedures set for receiving, grouping/categorizing, and recording all incoming information received by an intelligence centre;

   b. **Evaluation** “appraises each item of information in respect of the reliability of the source and the credibility of the information” (CFJP, 2011, p. 3-7). In other words, it is an assessment of the reliability of the source and the validity of the information originating from the source. Therefore, a rating is allocated to each piece of information or intelligence as a means to indicate the degree of confidence placed upon it;

   c. **Analysis** is the step “in which processed information is reviewed in order to identify significant facts for subsequent interpretation” (CFJP, 2011, p. 3-9). In analysis, the collated and evaluated information is scanned for significant facts and related to already known facts. Subsequently, deductions are formulated based on the comparison;

   d. **Integration** “enables the creation of a coherent intelligence picture through the synthesis of deductions drawn by analysis” (CFJP, 2011, p. 3-9). The synthesis of analyzed information gathered from a variety of sources allows the analyst to recognize meaningful patterns and relationships. This fusion could be a large amount of information or it could be a small amount of data integrated into data that has already been analyzed an integrated. Ideally, the information should be presented visually as this allows the analyst to better see and recognize the meaningful patterns and relationships.

e. **Interpretation** is the step "in which the significance of integrated information and intelligence is judged in relation to the commander's mission information requests, and basic intelligence, to create finished intelligence" (CFJP, 2011, p. 3-10). Interpretation is the cognitive process of comparison and deduction based on common sense, past experiences, knowledge of adversarial and friendly forces, and available information and intelligence. New information is compared with, or added to, with existing information which results in fresh intelligence.

4.2 Categorization of Open Source Information

As part of the classification schema, open source information was categorized into the following groups:

1. **Media:** This is mass media or news media that focuses on delivering news to the general public or a target public. Examples include online newspapers, magazines, and RSS feeds.

2. **Web-based communities and user generated content**: This group can be defined as websites that allow users to interact and collaborate with one another by generating content in a virtual community. These sites are in contrast to ones where people are limited to passively viewing of content. This category contains social networking sites, video-sharing sites, wikis, and blogs.

3. **Public data.** This is data that is readily available to the general population through the Internet such as government reports, budgets, government hearings and contract awards.

4. **Legal.** This group includes law enforcement data, legal documents, and court proceedings.

5. **Geospatial.** This is information with geospatial dimensions and includes hard and soft copies of maps and atlases, gazettes (public journals or newspapers of record), port plans, gravity data, aeronautical data, navigation data, geodetic data (earth measurements), environmental data and commercial imagery.

4.3 Software Tools Groupings

The categorization of the software tools presented in Section 4 is displayed in Table 4-1. For the first four categories of open source information (i.e., media, web-based, public data, and legal), each table cell has been split into 21 sub-cells. Each sub-cell is dedicated to a software tool as depicted below. The intent is to allow the reader to quickly scan the classification scheme to ascertain those tools that support multiple classes of open source information and/or multiple stages of the intelligence cycle.

EMM OSINT	Maltego	Wynyard
Recorded Future	Palantir	Addictomatic
Kapow	Mario's	Webcase
Rosette	ORA/Automap	NewsNow
Silobreaker	Social Mention	Who's Talkin
Kurrently	Samepoint	Hootsuite
Trackur	Marketing Cloud	

The software tools tailored towards handling of geospatial information do not overlap with the first four categories of open source information. As such, the sub-cells for this class of information are divided into 6 sub-cell as seen below.

Google Earth	Esri ArcGIS	PCI Geomatics
GRASS	Intergraph	Jagwire

The intent of this activity is to demonstrate that there is a capacity to search, import, categorize, and analyze information using those tools identified. Moreover, it provides a first pass at understanding the current landscape of available tools for the purposes of supporting the collection and collation (and analysis, as a secondary objective) of open source information.

Table 4-1: Categorization of Open Source Intelligence Tools

	COLLECTION			PROCESSING					
				Collation			Analysis		
MEDIA									
Online Newspapers	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard
	Recorded Future	Palantir	Addictomatic	Recorded Future	Palantir	Addictomatic	Recorded Future	Palantir	
	Kapow	Mario's	Webcase	Kapow	Mario's	Webcase			
	Rosette		NewsNow	Rosette		NewsNow			
Online Magazines	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard
	Recorded Future	Palantir	Addictomatic	Recorded Future	Palantir	Addictomatic	Recorded Future	Palantir	
	Kapow	Mario's	Webcase	Kapow	Mario's	Webcase			
	Rosette		NewsNow	Rosette		NewsNow			
Computer-Based Information	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard
	Recorded Future	Palantir	Addictomatic	Recorded Future	Palantir	Addictomatic	Recorded Future	Palantir	
	Kapow	Mario's	Webcase	Kapow	Mario's	Webcase			
	Rosette			Rosette					
RSS Feeds	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard
	Recorded Future		Addictomatic	Recorded Future		Addictomatic	Recorded Future		
	Kapow	Mario's		Kapow	Mario's				
	Rosette			Rosette					
WEB-BASED & USER-GENERATED									
Social Networking	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard		Maltego	Wynyard
	Recorded Future		Addictomatic	Recorded Future		Addictomatic	Recorded Future		

	COLLECTION			PROCESSING					
				Collation			Analysis		
	Kapow	Mario’s	Webcase	Kapow	Mario’s	Webcase			
	Rosette	ORA/Automap		Rosette	ORA/Automap			ORA/Automap	
	Silobreaker	Social Mention	Who’s Talkin	Silobreaker	Social Mention	Who’s Talkin	Silobreaker		
	Kurrently	Samepoint	Hootsuite	Kurrently	Samepoint	Hootsuite		Samepoint	Hootsuite
	Trackur	Marketing Cloud		Trackur	Marketing Cloud			Marketing Cloud	
			Addictomatic			Addictomatic			
			Webcase			Webcase			
		Samepoint			Samepoint				
	Trackur			Trackur					
Wikis	No tools								
	Recorded Future			Recorded Future			Recorded Future		
	Kapow			Kapow					
	Rosette		NewsNow	Rosette		NewsNow			
		Samepoint			Samepoint			Samepoint	
	Trackur			Trackur					
PUBLIC DATA									
Government Reports	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard	EMM OSINT	Maltego	Wynyard
	Recorded Future			Recorded Future			Recorded Future		
			NewsNow			NewsNow			NewsNow
Budgets/Hearings/ Contracts			Wynyard			Wynyard			Wynyard
LEGAL									

	COLLECTION			PROCESSING					
				Collation			Analysis		
Legal Intelligence			Wynyard			Wynyard			Wynyard
		Palantir			Palantir			Palantir	

	COLLECTION			PROCESSING					
				Collation			Analysis		
GEOSPATIAL									
Maps	Google Earth	Esri ArcGIS	PCI Geomatics	Google Earth	Esri ArcGIS	PCI Geomatics		Esri ArcGIS	PCI Geomatics
	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire
Atlases	Google Earth			Google Earth			Google Earth		
Gazettes	Google Earth			Google Earth					
Port Plans	Google Earth			Google Earth					
Gravity Data	No tools								
Aeronautical Data	Google Earth			Google Earth					
Navigation Data	Google Earth	Esri ArcGIS	PCI Geomatics	Google Earth	Esri ArcGIS	PCI Geomatics		Esri ArcGIS	PCI Geomatics
	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire

	COLLECTION			PROCESSING					
				Collation			Analysis		
Satellite Data	Google Earth	Esri ArcGIS	PCI Geomatics	Google Earth	Esri ArcGIS	PCI Geomatics		Esri ArcGIS	PCI Geomatics
	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire
		Intergraph			Intergraph			Intergraph	
Environmental Data	Google Earth	Esri ArcGIS	PCI Geomatics		Esri ArcGIS	PCI Geomatics		Esri ArcGIS	PCI Geomatics
	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire	GRASS	Intergraph	Jagwire
Commercial Imagery	Google Earth			Google Earth					

5 Conclusions and Recommendations

A search of open source resources was completed in order to understand the landscape with respect to the software tools that are readily available to support the automated collection and collation of open source information. The intent was to create an understanding of potential technological options for supporting OSINT activities within CJOC. To that end, this preliminary search revealed the following high-level findings:

- There are numerous existing tools with a select number available free of charge to support the collection and collation of OSINT.

- Individual tools typically provide functionality to support the several phases of the Intelligence cycle (i.e., collection, collation, and analysis).

- Individual tools are generally tailored to handle a specific class of OSINT (i.e., media, geospatial); however, certain tools possess functionality to handle multiple classes of OSINT material.

As next steps, the following recommendations are put forth:

- Conduct data collection activities, including interviews with CJOC personnel, in order to understand and document the current OSINT processes and tools, storage methods and refresh intervals for OSINT.

- Define the gaps and bottlenecks with respect to the execution of current CJOC processes for collecting and collating OSINT.

- Complete a ‘fit’ analysis to determine the feasibility of implementing software tools identified in this report to address the gaps and bottlenecks with the current CJOC OSINT processes. The ‘fit’ should look at both the operational perspective (e.g., improving search capability) and technical perspective (e.g., compatibility with existing hardware, interoperability with existing OSINT applications, certification and accreditation).

1. References

This section identifies all papers, proceedings and book chapters that are included in Sections 3.1 and 3.2. The references are in American Psychological Association (APA) format and are listed in ascending order by date (year) and within each year, they are listed alphabetically.

Reid, E., Qin, J., Chung, W., Xu, J., Zhou, Y., Schumaker, R., & Chen, H. (2004). Terrorism knowledge discovery project: A knowledge discovery approach to addressing the threats of terrorism. In H. Chen, R. Moore D. Zeng & J. Leavitt (Eds.), *Lecture Note in Computer Science: Vol. 3073. Intelligence and Security Informatics*, 125-145. doi:10.1007/978-3-540-25952-7_10

Carroll, J. M. (2005). OSINT analysis using adaptive resonance theory for counterterrorism warnings. *Artificial Intelligence and Applications,* 756-760.

Koltuksuz, A., & Tekir, S. (2006*).* Intelligence analysis modeling. In N. Szczuka, D. Howard, D. Slezak, H. Kim, T. Kim. I. Ko, G. Lee & P. Sloot (Eds.), *Advances in Hybrid Information Technology, 1,* 146-151. doi:10.1109/ICHIT.2006.253479

Neri, F., & Baldini, N. A. (2006, October). *Multilingual Text Mining based content gathering system for Open Source Intelligence*. Paper presented at the International Atomic Energy Agency Conference, Wien, Austria.

Badia, A., Ravishankar, J., & Muezzinoglu, T. (2007). Text Extraction of Spatial and Temporal Information*. Proceedings of the Intelligence and Security Informatics Conference*, *USA,* 381. doi:10.1109/ISI.2007.379527.

Baldini, N. Neri, F., & Pettoni (2007*). A multilanguage platform for open source intelligence*. In A. Zanasi, C. Brebbia & N. Ebecken (Eds.), *Data Mining VIII: Data, Text and Web Mining and their Business Applications: Vol. 38* (pp. 18-20). New Forest, UK. doi:10.2495/DATA070321

Hopewell, P. H. (2007). *Assessing the acceptance and functional value of the Asymmetrical Software Kit (ASK) at the Tactical Level*. (Unpublished Master's Thesis). Naval Postgraduate School, Monterey, California.

Memon, N., Hicks, D. L., & Larsen, H. L. (2007). Harvesting Terrorists Information from Web. *Proceedings of the 11th International Conference on Information Visualization, Switzerland, IV07,* 664-671. doi: 10.1109/IV.2007.60

Thibault, G., Gareau, L. M., & Le May, F. (2007). Intelligence collation in asymmetric conflict: A Canadian armed forces perspective. *Proceedings of the 10th Annual Conference on Information Fusion, Quebec City, Quebec*. doi: 10.1109/ICIF.2007.4408115

Ulicny, B., Baclawski, K., & Magnus, A. (2007). New metrics for blog mining. In N. Glance, N. Nicolov, E. Adar, M. Hurst, & F. Salvettii (Eds.), *Proceedings of the 1st*

International Conference on Weblogs and Social Media, Boulder, CO, USA. doi:10.1117/12.720454

Zanasi, A. (2007). *New forms of war, new forms of intelligence: Text mining.* Paper presented at the Information Technology for National Security Conference, Riyadh, Saudia Arabia.

Best, C. (2008). Web mining for open source intelligence*. Proceedings of the 12th International Conference on Information Visualization*, *England*, *IV08*, 321-325. doi:10.1109/IV.2008.86

Kallurkar, S. (2008). *Targeted information dissemination.* (Report No. Unknown). Retrieved from the Defense Technical Information Center (DTIC) Website: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA480150.

Neri, F. & Pettoni, M. (2008) *Stalker, a multilingual text mining search engine for open source intelligence. Proceedings of the 12th International Conference on Information Visualization*, *England, IV08,* 314-320. doi:10.1109/IV.2008.9

Pfeiffer, M., Avila, M., Backfried, G., Pfannerer, N., & Riedler, J. (2008). Next Generation Data Fusion Open Source Intelligence (OSINT) System Based on MPEG7. *Proceedings of the Conference on Technologies for Homeland Security USA,* 41-46. doi:10.1109/THS.2008.4534420

Fei, Z., Xu, H., Weisheng, X., & Qidi, W. (2009). Analysis and Design of Web-Based Intelligence Mining Service System. *Proceedings of the Management and Service Science Conference*, *USA,* 1-4. doi:10.1109/ICMSS.2009.5300887

Katakis, I., Tsoumakas, G., Banos, E, Bassiliades, N., & Vlahavas, I. (2009). An adaptive personalized news dissemination system. *Journal of Intelligent Information Systems: 32,* 191-212. doi: 10.1007/s10844-008-0053-8

Neri. F, & Geraci, P. (2009). Mining textual data to boost information access in OSINT. *Proceedings of the 13th Conference on International Information Visualization, Spain, IV09,* 427-432. doi: 10.1109/IV.2009.99

Pouchard, L. C., Dobson, J. M., and Trien, J. P. (2009, March). *A framework for the systematic collection of open source intelligence*. Paper presented at the meeting of the Association for the Advancement of Artificial Intelligence Conference, Palo Alto, CA, USA.

Neri, F., Geraci, P., & Camillo, F. (2010*). Monitoring the Web Sentiment, The Italian Prime Minister's Case. Proceedings on the Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference, Denmark,* 432-434. doi:10.1109/ASONAM.2010.26

Boury-Brisset, A. C., Frini, A., & Lebrun, R. (2011, June). *All-source Information Management and Integration for Improved Collective Intelligence Production*. Paper presented at the 16th Annual International Command and Control Research and

Technology Symposium – Collective C2 in Multinational Civil-Military Operations, Quebec City.

Gibson, S. D. (2011). *Open source intelligence (OSINT): A contemporary intelligence lifeline.* (Doctoral dissertation). Retrieved from Cranfield Defence and Security, Shrvenham database (1826/6524).

Piskorski, J., Tanev, H., Atkinson, M., van der Goot, E., & Zavarella, V. (2011). Online news event extraction for global crisis surveillance. In N. Nguyen (Ed.), *Lecture notes in Computer Science: Vol. 6910: Transactions on Computational Collective Intelligence,* 182-212. doi:10.1007/978-3-642-24016-4_10

Qureshi, P. A. R., Memon, N., Wiil, U. K., Karampelas, P., & Sancheze, J. I. N. (2011). Harvesting Information from Heterogeneous Sources. *Proceedings from the European Conference on Intelligence and Security Informatics, Greece,* 123-128*.* doi:10.1109.EISIC.2011.76

Roberts, N. C. (2011). Tracking and disrupting dark networks: Challenges of data collection and analysis. *Information Systems Frontiers, 13(1),* 5-19. doi: 10.1007/s10796-010-9271-z

Roy, J., & Auger, A. (2011, June). *The multi-intelligence tools suite – Supporting research and development in information and knowledge exploitation.* Paper presented at the 16th International Command and Control Research and Technology Symposium – Collective C2 in Multinational Civil-Military Operations, Quebec City, Canada.

Noubours, S., & Hecking, M. (2012). Automatic exploitation of multilingual information for military intelligence purposes. *Proceedings of the Military Communications and Information Systems Conference (MCC), Poland*, 1-8.

Ríos, S. A., & Muñoz, R. (2012). Dark Web portal overlapping community detection based on topic models. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, *2.* doi: 10.1145/2331791.2331793

Su, P., Li, D., & Su, K. (2012). An expected utility-based approach for mining action rules. *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*, *China.* doi:10.1145/2331791.2331800

Yang, H. C., & Lee, C. H. (2012, August). Mining open source text documents for intelligence gathering. In X. Jiang (Chair), *International Symposium on Technology in Medicine and Education*. Symposium conducted at the IEEE Sapporo Section, Hokkaido, Japan.

This page intentionally left blank.

List of symbols/abbreviations/acronyms/initialisms

ANN	Artificial Neural Network
APA	American Psychological Association
API	Application Program Interface
ASIC	All Source Intelligence Center
ASK	Asymmetric Software Kit, Asymmetrical Software Kit
CANDID	Canadian Defence Information Database
CASOS	Computational Analysis of Social and Organizational Systems
CFJP	Canadian Forces Joint Publication
CJOC	Canadian Joint Operations Command
CORA	Centre for Operational Research and Analysis
COIC	Commandement des opérations interarmées du Canada
CSV	Comma Separated Value
DGInet	Distributed Geospatial Network
DRDC	Defence Research and Development Canada
DTIC	Defense Technical Information Center
EMM	Europe Media Monitor
GHz	gigahertz
GIS	Geographic Information System
GPS	Global Positioning System
GRASS	Geographic Resources Analysis Support System
HTML	Hyper Text Markup Language
HUMINT	Human Intelligence
ICT	Information and Communication Technology
IDM	Investigative Data Mining
IMINT	Imagery Intelligence
IMSS	Intelligence Mining Service System
IP	Internet Protocol
IPB	Intelligence Preparation of the Battlefield
ISR	Intelligence, Surveillance, and Reconnaissance
JMS	Java Message Service
JOIC	Joint Operation and Intelligence Center
JSON	JavaScript Object Notation
KHz	kilohertz
LiDAR	Light Detection and Ranging
MHz	megahertz
NATO	North Atlantic Treaty Organisation
NLP	Natural Language Processing
OSINT	Open Source Intelligence
QLI	Quantum Leap Innovations
RAND	Research and Development
RDDC	recherche et de développement pour la défense
RGB	Red, Green, Blue
RSS	Rich Site Summary
RTO	Research and Technology Organisation

SIGINT	Signals Intelligence
SIG	système d’information géographique
SMS	Short Message Services
SMT	Statistical Machine Translation
SOA	Services Oriented Architecture
SOAP	Symbolic Optimal Assembly Program
SQL	Structured Query Language
TID	Targeted Information Dissemination
URL	Uniform Resource Locator
USASOC	U.S Army Special Operations Command
XML	Extensible Markup Language

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a
contractor's report, or tasking agency, are entered in section 8.)

Lisa Hagen, CAE Professional Services

2. SECURITY CLASSIFICATION
(Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED
(NON-CONTROLLED GOODS)
DMC A REVIEW: GCEC JUNE 2010

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U)
in parentheses after the title.)

Methods and Tools for Automated Data Collection and Collation of Open Source Information

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

Hagen, L.

5. DATE OF PUBLICATION
(Month and year of publication of document.)

August 2013

6a. NO. OF PAGES
(Total containing information, including Annexes, Appendices, etc.)

52

6b. NO. OF REFS
(Total cited in document.)

30

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Contract Report

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

Defence R&D Canada – CORA
Dept. of National Defence, MGen G. R. Pearkes Bldg., 101 Colonel By Drive, Ottawa
ON K1A 0K2, CanadaDefence R&D Canada – CORA
101 Colonel By Drive
Ottawa, Ontario K1A 0K2

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document
was written. Please specify whether project or grant.)

N/A

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

W7714-083663/001/SV

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

DRDC CORA CR 2013-119

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

Unlimited

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

A search of open source resources was completed in order to understand the landscape with respect to the software tools that are readily available to support the automated collection and collation of open source information. The intent is to create an understanding of potential technological options for supporting open source intelligence (OSINT) activities within the Canadian Joint Operational Command (CJOC). To that end, this preliminary search revealed the following high-level findings:

- There are numerous existing tools with a select number available free of charge to support the collection and collation of OSINT.

- Individual tools typically provide functionality to support the several phases of the Intelligence cycle (i.e., collection, collation, and analysis).

Individual tools are generally tailored to handle a specific class of OSINT (i.e., media, geospatial); however, certain tools possess functionality to handle multiple classes of OSINT material.

Un examen des ressources ouvertes a été effectué dans le but de comprendre la situation en ce qui concerne les logiciels disponibles pour soutenir la collecte et le regroupement automatiques de l’information de sources ouvertes. L’objectif est d’assurer une compréhension des possibilités technologiques visant à soutenir les activités du renseignement de sources ouvertes (OSINT) au sein du Commandement des opérations interarmées du Canada (COIC). À cette fin, cet examen préliminaire en est arrivé aux conclusions de haut niveau suivantes :

- Il existe de nombreux outils, dont un certain nombre sont gratuits, permettant de collecter et de regrouper l’OSINT.

- Les outils offrent généralement des fonctions visant à soutenir les différentes phases du cycle du renseignement (collecte, regroupement et analyse).

- Les outils sont généralement adaptés au traitement d’une catégorie précise d’OSINT (médiatiques ou géospatiaux, par exemple), mais certains outils sont en mesure de traiter plusieurs classes d’OSINT.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Collection, Collation, Intelligence, Open Source, OSINT, Automated