Semantic Technology:
Intelligent Solutions for Big Data Challenges

Foreword

The study presents a joint project of Advanced Analytics, experts in the field of semantic technology, and EXOP, a German risk consultancy and analysis firm. It is intended to provide a summary of recent trends in the Internet that form a nexus between security affairs and the latest developments in information technology: OSINT, SOCMINT, Big Data Analysis and semantic technologies have all become essential aspects in the day-to-day work of analysts. Collecting data, identifying risks and threats, analysing trends: this study attempts to give an overview of the contributions semantic technology intelligence platforms can offer to a variety of fields, above all to the intelligence cycle.

Authors List

Nicole Stracke is Director of Advanced Analytics GmbH. Previously, she acted as Director of the Centre for Research & Analysis with EXOP GmbH, a German risk management consultancy. Ms Stracke has long-standing experience working as a consultant and lead analyst for the Middle East and North Africa region. In the course of her professional career, she has spent several years in the Middle East and has conducted projects in several crisis regions, such as Yemen, Sudan, Iraq and Syria.

Matteo Grella, Associate Director of Advanced Analytics GmbH, is the executive leader of the Semantic Software development process. Mr Grella has worked as an expert and consultant on Knowledge Management, Media Monitoring and OSINT with private companies (e.g. Alkemy, RESI, KPMG, GroupM) and academic research institutes, including the University of Turin, the Politecnico di Torino and the ILC-CNR of Pisa. Matteo is a member of the AI*IA and the Association „Senso Comune“ -Rome, and an author of scientific publications in his area of expertise.

Dr. Bernhard Chiari, Director Organisations & Government, is responsible for EXOP's public sector clients. He received a PhD in Eastern European History from Tübingen University and worked in Minsk, Moscow, St. Petersburg and at Erlangen University. Before joining the EXOP Group, Bernhard held various executive positions in the Centre of Military History and Social Sciences of the Bundeswehr (ZMSBw). As a reserve staff officer in the military intelligence branch, he served several times on the Balkans and in Afghanistan.

Karsten Bechle is Associate Director of the Centre for Research & Analysis with EXOP GmbH. He is responsible for all research and analysis carried out by EXOP's Americas and Europe/Community of Independent States departments. Before joining EXOP, he worked as a researcher at the GIGA German Institute of Global and Area Studies in Hamburg and at the Arnold-Bergstraesser-Institute/University of Freiburg. He has conducted intense field and desk research on Latin America and has vast on-the-ground experience in numerous countries of the region.

Stephan Schmid is Associate Director of the Centre for Research & Analysis with EXOP GmbH, and is responsible for all research and analysis carried out by EXOP's Middle East & North Africa and Sub-Saharan Africa departments. Before joining EXOP as a senior risk analyst, Mr Schmid worked in various research institutions in the Middle East and Germany. Mr Schmid has lived in and travelled extensively throughout the MENA region, including Lebanon, Syria, Jordan and Israel/Palestine.

Contents

Executive Summary

Executive Summary

Over the past decade, social and open-source media have become more and more dominant in global communication, interaction and analysis. The ever-growing volumes of both unstructured and structured data on the Web have slowly given rise to a pervasive feeling of being faced with a constant information overload. The sheer abundance of data is a challenge for decision makers, who often have to react to critical situations quickly. When the time to acquire the knowledge necessary for a profound understanding of an issue is not at hand, the immediate orientation within available data floods is a critical precondition for success.

The current trend – the digitalisation of all data – will ultimately lead to even larger amounts of information, creating an urgent necessity for solutions able to support:

- the analysis of structured and unstructured data;
- the visualisation of data to recognise new information and causal relationships;
- the rapid identification and selection of essential information and its dissemination to selected recipients in real time;
- the integration of all internal and selected external databases into a single knowledge architecture that can be administered easily (flexible sources, keywords, concepts, user management, etc.).

Leading enterprises have been investing heavily in research programmes focusing on artificial intelligence and semantic technology. Today, semantic technology is being applied across various sectors, such as health care, communications, transport, logistics, security and defence. However, there is still significant scope for development. Criticism of existing search and intelligence platforms based on semantic technology has mainly centred on their inefficiency and lack of accuracy, as well as the difficulties of integrating such platforms into the IT infrastructure of companies and organisations. Furthermore, maintaining and updating platforms often comes at high costs while providing only a low degree of flexibility for the users.

As the digital revolution continues relentlessly, the need for artificial intelligence platforms is increasing. New technologies are being developed in the field of Natural Language Processing. This technology already yields decisive benefits today and will do so even more in the upcoming years. This is especially true for all forms of intelligence and analysis. As shown in this study, the application of newly developed semantic technologies brings clear advantages at every step of the intelligence process cycle, with a focus on:

- streamlining and simplifying the workflow of gathering, evaluating, analysing and disseminating data/information;
- facilitating the identification and analysis of risks and threats;
- automatically adapting to specific and changing requirements of users due to artificial intelligence components, thereby supporting the intelligence process.

Over the past year, semantic technology has evolved into a powerful tool for the analysis of risks and threats. Particularly the security and intelligence sectors can gain tremendous benefits from the advantages that the latest devlopments in semantic technology are able to provide.

Web 2.0 as a Platform and Space for Global Communication, Interaction and Analysis

Web 2.0 as a Platform and Space for Global Communication, Interaction and Analysis

Social Media: A Tool for Global Jihadist Activities

Jihadists around the world use social media to spread their propaganda, attract recruits and issue threats. Social media accounts are opened and abandoned rapidly and diversified according to the structures, cultures and languages in the respective countries of origin; news are employed to distort information. This complex environment poses a challenge to the efforts of intelligence services and analysts to detect and evaluate threats in time.

Current Developments

Currently, 40,000–50,000 jihadist social media accounts exist worldwide, each of them posting several messages per day.

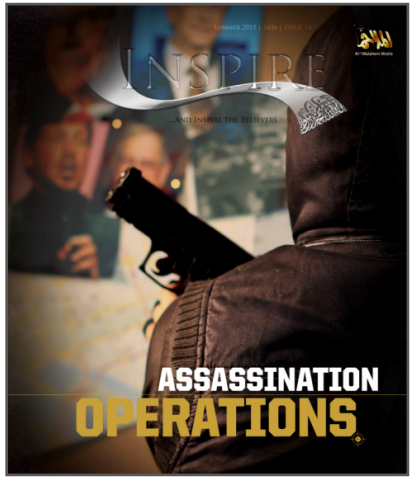
The jihadist social media network creates an impressive amount of information. It is an integral element of and an important tool for the worldwide jihadist community that facilitates recruitment and participation.

Being part of the jihadist social media community is the first step of becoming active in real-world jihadist activities. The opportunities afforded by social media have to a great degree helped to obtain funding, provide services, carry out lone-wolf attacks and recruit members through religious instruction.

Al-Qaida in the Arabian Peninsula – The Early Phase of Using Web 2.0

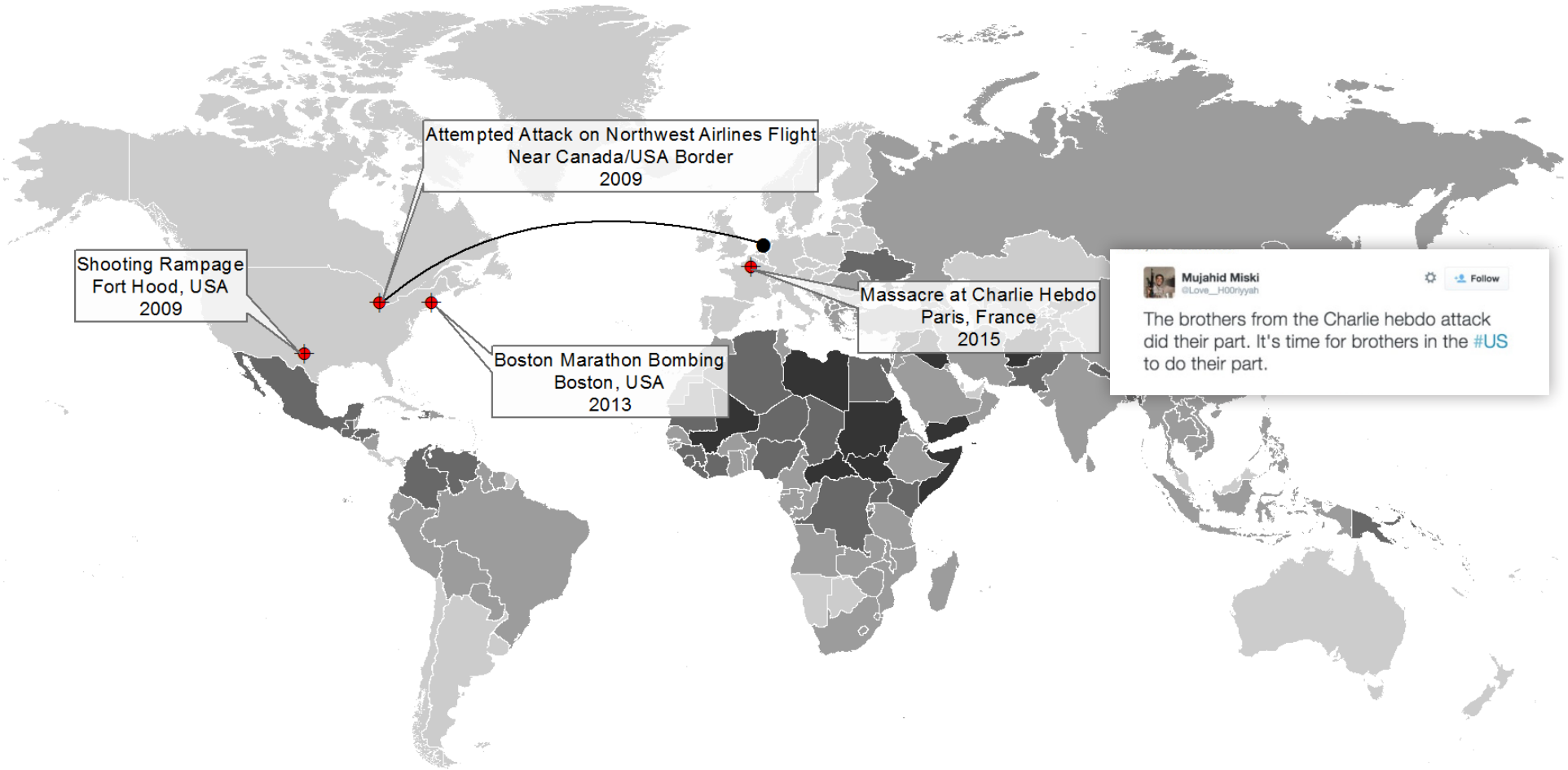
The Yemeni branch of al-Qaida (al-Qaida in the Arabian Peninsula – AQAP) was a pioneer among jihadist groups in using the Internet for Islamist purposes, in particular propaganda activities. In 2009, the group published and distributed the first issue of its bi-monthly Arabic-language online magazine *Sada al-Malahim* (Echo of Battles) via blogs and forums on the Internet. One year later in July 2010, the group recognised the importance of publishing propaganda in English in order to reach a global audience. The first English publication appeared under the title *Inspire*, which highlighted three main objectives of AQAP's media strategy:

- ideological indoctrination of non-Arabic-speaking partisans and potential supporters in Western countries;
- practical instructions for making bombs/ detailed descriptions of strategies for carrying out attacks in the West;
- inciting radical-minded Muslims living in the West to carry out independent attacks.



While the English magazine addresses Western Muslims, the Arabic version *Sada al-Malahim* is predominantly designed to influence the mindset of the local Yemeni and Saudi Arabian population. It focuses on spreading AQAP's ideological tenets and recruiting new members by presenting jihad as a national struggle and a solution to local grievances. As a result, AQAP propaganda strategies on the internet and particularly in social media have directly inspired and instigated manifold terrorist operations against the West.[1]

AQAP - inspired attacks



IS – Mastering Social Media Propaganda

Going a step further, the Islamic State (IS) redefined the methods of internet usage and brought them to an entirely new level. IS extensively addresses both Muslims and non-Muslims around the world via social media. Individual IS supporters try to influence their communities through multi-lingual accounts on various platforms. These activists know how and with which topics to attract maximum attention in their respective environment.

Consequently, tweets and retweets have been spreading IS ideology and propaganda in an unprecedented speed. The organisation has been able to create efficient and sustainable support networks and to put its objectives into action in remarkably short periods of time. [2]

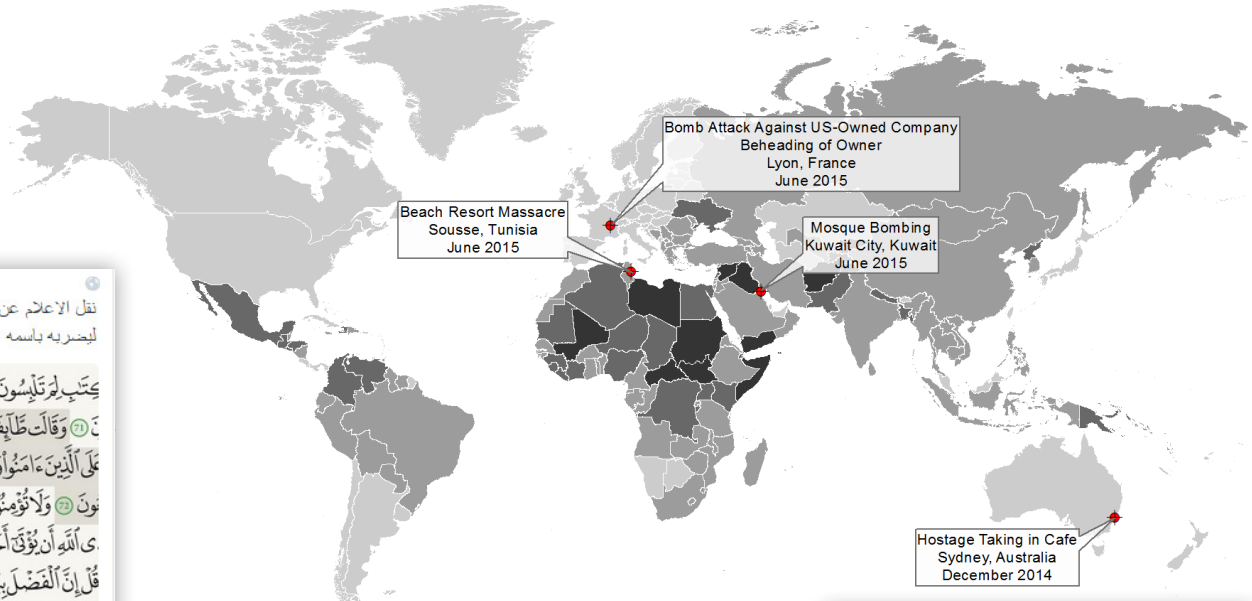
A large number of tweets and high-profile announcements by the IS leadership and affiliates have successfully encouraged like-minded jihadists to stage attacks worldwide[3]. Such incidents are subsequently discussed and commented on extensively in the jihadist social media network to incite further similar attacks.

Example for social media platforms predominantly used by IS

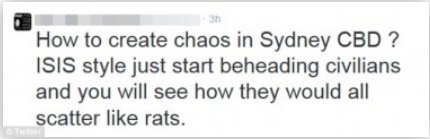
Purposes of extensive social media usage by IS

Example for IS inspired attacks

IS - inspired attacks



Tweet of an IS supporter announcing that the „first Israeli Jew“ has joined the ranks of IS



Comments by Wilayat Sina', the Egyptian branch of IS, on recent developments concerning IS

Challenges and Deficits

The increasing use of social media among militant groups for spreading propaganda, recruitment and announcing attacks presents a critical security challenge for intelligence services and other authorities and enterprises. When monitoring militant groups, civil and military analysts face a number of difficulties. They have to:

- follow an increasing number of social media accounts and media outlets;
- keep an eye on which accounts have been opened or abandoned;
- identify distorted/false information provided by social media or news outlets;
- analyse information in an increasing number of different languages;
- keep abreast of the emergence of new Social Media platforms that are used by militant groups;
- analyse the changing attitude (negative, positive, neutral) of a group in regard to a topic or political development.

IS threatens to move its activities into Algeria

In order to tackle these challenges and to facilitate the process of intelligence gathering, evaluation and threat validation, enterprises and state authorities have started to implement software solutions of various kinds. These applications are intended to support the analysts' work process. Some of the basic requirements of such applications are:

- the option of real-time research in a single data pool that combines various types of sources, including SOCMINT, broadcasting and Open Source Intelligence (OSINT);
- automated identification of newly emerged groups as well as abandoned or newly opened accounts that are linked to the specific subject of monitoring;
- the exploitation of sources in different languages without delays caused by human translation;
- learning analysis functionalities, such as the automatic connection of names, relations, opinions and locations, supported by visualisation tools;
- intelligent software support for the detection of keywords and entities in sources in different languages, and for the identification and prediction of trends;
- the option of integrating and analysing unstructured data (social media, video and audio files) for information evaluation and analysis.

The use of social media among militant groups for spreading propaganda and recruitment presents a critical challenge for analysts.

Averting Security Risks by Using Social Networks: A Case Study from Mexico

Over the past several years, social media such as Facebook, Twitter and various blogs have become the most relevant source of information. This is also the case when it comes to the early detection and avoidance of violent conflicts with a high potential of affecting innocent bystanders.

Security Situation in Mexico

Since 2006, over 70,000 people have been killed in the Mexican drug war[4]. In some parts of the country, shoot-outs between members of rivalling drug cartels or between criminals and security forces have become a regular occurrence. Sometimes confrontations take place in broad daylight on major toll roads or in city centres. Blocking roads with burning vehicles is a common means of both deterring security forces from arriving at a crime scene and retaliating against arrests or killings of major kingpins.

Drug-related violence generally erupts spontaneously and without prior warning. The fear of being in the wrong place at the wrong time is widespread among citizens. As a result, many Mexicans in the areas most affected by the drug war have adjusted their everyday behaviour accordingly. One way that people have adapted to the deteriorating security situation is by not travelling or leaving their house after dark. No less important, however, is the turn to social media as a means of informing themselves and others about potential security risks.

Facebook site dedicated to reporting on crime in and around Torreón

The Role of Social Media

Social media have become a vital source of information for many Mexicans worried about security[5]. The unstable security environment in many parts of the country has created an increased demand for quick and unrestricted access to information on security-relevant developments. This need has been met by a mushrooming number of blogs, tweets and Facebook accounts dedicated to reporting on crime and violence in specific parts of the country. In the metropolitan areas of such cities as Tampico, Reynosa, Torreón, Monterrey and Veracruz, it is fairly common to check the security situation with the help of social media before setting out at night. Moreover, citizens frequently not only use social media to inform themselves about security-critical developments, but also actively report security incidents and potential risks via these channels.

> The increased demand for quick and unrestricted access to information on security relevant developments in Mexico has been met by a mushrooming number of blogs, tweets and Facebook accounts dedicated to reporting on crime and violence.

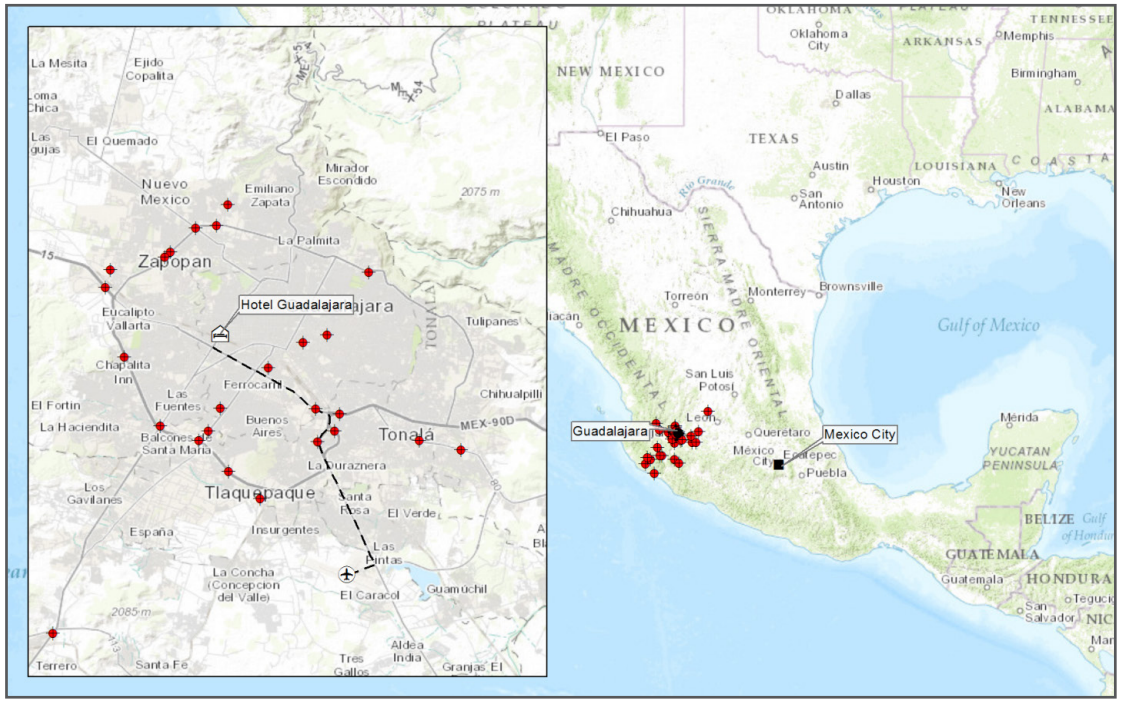
The reasons for the outstanding role of social media are manifold. For one, many local Mexican newspapers refrain from publishing incidents related to the country's drug war. Numerous reporters covering drug-related crimes have been killed, leaving editorial staff with little choice but to resort to a rising degree of self-censorship in order to protect themselves. Blogs and other social media have come to fill this void. In addition, tweets and posts are generally disseminated earlier and spread faster than news in the traditional media, a crucial factor when it comes to avoiding security threats.

Case Study Guadalajara

On 01 May 2015, members of a Mexican drug cartel blocked dozens of roads in the state of Jalisco in a concerted action, virtually closing all access ways to the capital Guadalajara, generally a rather calm city. The coordinated attacks – which also led to the downing of a military helicopter with a rocket launcher in a remoter area of Jalisco – was a retaliatory measure by a local drug cartel against the earlier arrest of one of its high-ranking leaders.[6]

Immediately after the clashes had started, posts and tweets with the hashtag "#narcobloqueo" (a neologism denoting roadblocks erected by drug cartels) reported on the events. Written comments were accompanied by pictures of burning trucks, buses, tank cars and other vehicles, giving a live impression of the scope of the attacks. The local security forces also made use of their twitter account to keep the population informed.

Tweets reporting roadblocks in Guadalajara



Roadblocks in Guadalajara and Jalisco State on 01 May 2015

Such events as the incidents in Guadalajara often come as a surprise, even to the local population. Even more, foreigners in the area would have been caught off guard. For example, a business traveller arriving at Guadalajara International Airport would have found it almost impossible to get from the airport to any hotel in the city centre without being forced to stop at one of the roadblocks.

Challenges and Deficits

The turn to social media as a means of informing oneself and others about potential security risks is an ongoing trend. A more systematic analysis of these information channels helps identify and avert threats and/or deal with specific incidents. Establishing an early-warning system based on SOCMINT/OSINT tools provides a substantial temporal advantage for proactively taking action or reacting quickly to possible threats and risks. Some of the basic requirements of such a system are:

- ensuring permanent 24/7 monitoring;
- real-time research in a single data pool that combines various types of sources (structured and unstructured data);
- sufficient capacities for 500 users or more to easily access and quickly distribute relevant information;
- having the option of an automatic and quick generation of alerts and standard reporting in the form of SMS, applications or e-mails;
- systemic learning analysis, such as the automatic connection of names, relations, opinions and locations, supported by visualisation tools.

Dealing With Information Overload: Orientation and Sense-Making in Complex Environments

The military's experience in Afghanistan has shown that understanding complex environments is key for stabilisation missions and state building. Capability gaps remain with regard to the analysis of manifold sources, including social media and social networks. Some of the shortfalls can be eliminated with the help of intelligent search tools and semantic software.

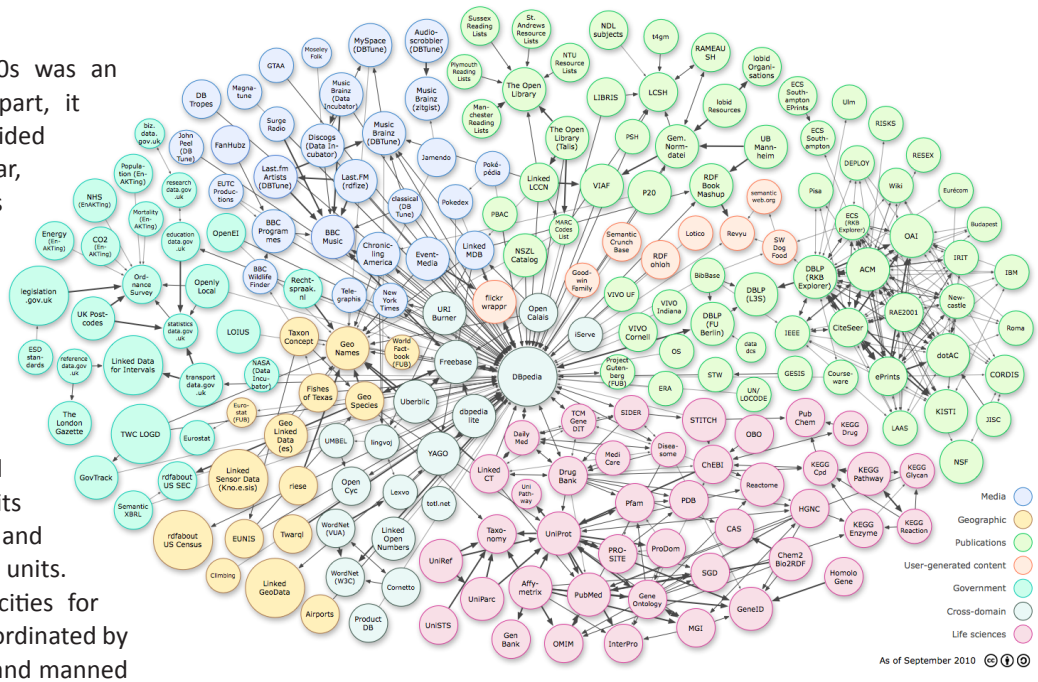
ISAF Mission: Understanding the Environment is Key

Until the end of the Cold War, intelligence dealt with comparably one-dimensional subjects. Intel officers analysed military structures, capacities and capabilities and tried to predict the course of enemy operations. In Afghanistan, civil-military stabilisation endeavours stood in the centre of all combined efforts. Operations took place in an environment that differed from conventional war scenarios with clearly distinguishable sides or parties. The mission of the International Security Assistance Force (ISAF) required the support of political, social and economic groups within the country. With regard to military intelligence, this approach demanded new professional skills. Analysts had to identify quickly changing enemy tactics, techniques and procedures. In addition, they were expected to fully understand the environment the Taliban acted in. Strictly military battlefield reconnaissance was replaced by the Intelligence Preparation of the Battlespace (IPB). The IPB theoretically covered all structures, to include political, religious and social groups, relations and conflicts, that were relevant to the combined military-civilian efforts. The populace's perception of the measures taken by the government and its Western supporters became a central object of intelligence work. Intel recommendations were intended to help military leaders make the right decisions in order to positively affect the social and economic system. Looking back on the Afghan experience, it goes without saying that this objective very often overburdened the responsible military and civilian personnel. [7]

Source: The Cyber & Jihad Lab

Variety of Military Sources

Conventional military intelligence in the 1990s was an almost closed system. To an overwhelming part, it relied on classified information often provided by national secret services. In the case of war, situational awareness on the battlespace was provided by reconnaissance units deployed as part of conventional forces. In Afghanistan, available sensors and sources were incomparably more manifold. Network Enabled Capabilities and Network Centric Warfare were concepts to cope with the local, more dynamic and less predictable character of the conflict. Military intel capabilities at different command levels comprised classical reconnaissance units with various technical sensors on the ground and additional intel personnel embedded in field units. Human intelligence teams, sensors and capacities for Electronic Warfare and Geoinformation were coordinated by respective staff cells. Unpiloted Aerial Vehicles and manned aircraft carried intel sensors of different kinds. Specialists for Civil-Military Cooperation, Information Operations and Cultural Advisory added their view of the situation. Secret services and other national reach-back capabilities, the different Special Operations Forces and the ANSF contributed to the situational picture. A flood of written reports and other products was the output of the daily work[8].



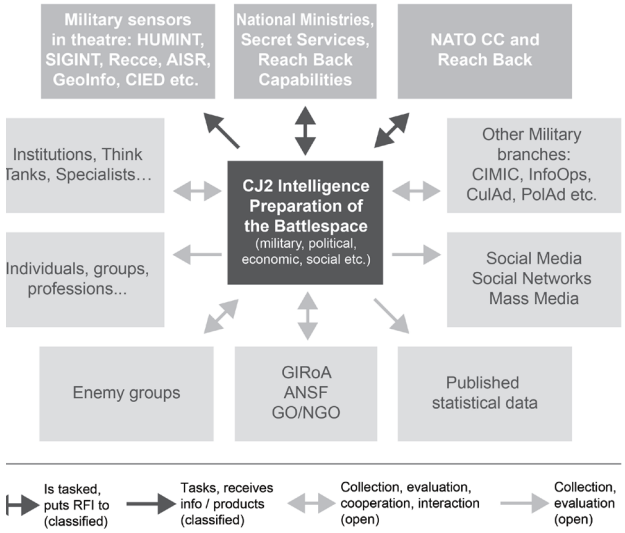
Source: Concordia University, Montreal

Sense-Making: Covering the Info Space

In order to face the complexity of the Afghan environment, however, the entire information space had to be taken into consideration. Press releases, statistical and economic data, transcriptions of radio talk shows and Shura-meetings reflected developments and sentiments in state and society. Chats and debates conducted in social media presented unique insights into Afghanistan's reality. Terrorist attacks and Afghan or Western countermeasures were intensively discussed in social networks. Governmental and Non-Governmental Organisations present in an area often assessed the local situation quite differently than the military. So did Afghan state officials or the media in Western countries. The Taliban and other opposing groups, finally, developed a high proficiency in using electronic media to support their campaigns. To get a realistic impression how things were developing in Afghanistan, all these forums had to be covered. [9]

Closing the Intel Cycle

The responsible analysts tried hard to fulfil Intel Collection Plans, to identify and categorise sources, to navigate a confusing landscape of data and reports and to close the Intel Cycle. Assessing the pertinent sources was a challenge due to technical reasons. Different information systems and networks (NATO and national, open and classified) were in place simultaneously. Analysts were confronted with the different languages spoken in Afghanistan and the neighbouring countries. The sheer amount of information made it difficult to filter and prioritise, to extract new insights and to identify trends. Provided search engines and analysis software helped to manually search for known keywords or to draw graphs and diagrams, but did not support the process of analysis itself. In practice, military intelligence often remained at the level of collecting highly detailed information on insurgents. Intel shops concentrated on the relationships between various groups and the possible location of enemy leaders. Intelligence therefore often failed to provide comprehensive analyses and vital general information on the environment in which the Taliban operated. Instead of trusting the recommendations of their intel specialists, decision makers sometimes consulted the mass media in order to reach the population and to effectively counteract the Taliban's influence. [10]



Challenges and Deficits

In order to master this challenge, the ISAF intel structures were enhanced continuously. Teams within so-called All Source Intelligence Cells did their best to fuse collected information and to produce comprehensive, operational intelligence. Multi-Fusion Centers were introduced to bring together military and civilian analysts. The military learned a lot from the Afghan experience with regard to structures, procedures and necessary country expertise. In order to deal with the described information overflow, however, several technical challenges could not be overcome until the end of the ISAF mission. Some of the basic requirements are:

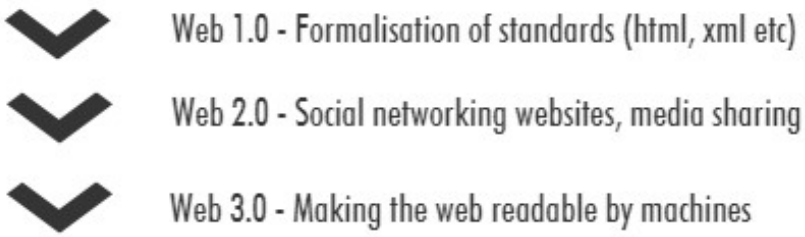
The sheer amount of information made it difficult to filter and prioritise, to extract new insights and to identify trends.

- real-time research in a single data pool that combines various types of sources;
- exploiting sources in different languages without delays by human translation;
- learning analysis functionalities, such as the automatic connection of names, relations, opinions and locations;
- intelligent software support for the detection of keywords and entities in sources in different languages, and for the identification and prediction of trends;
- integrated, user-friendly visualisation tools.

How Will We Interact With the Next-Generation Web?

How Will We Interact With the Next-Generation Web?

Semantic technology, Internet of Things, artificial intelligence, open technologies, intelligent Web: these catchwords indicate the paths the next-generation Web is taking. At the same time, they point to the question of what risks these new developments may hold for security analysis & intelligence.



Trajectory

The first generation of the Internet, also called Web 1.0, focused on protocols such as HTTP, on open standard markup languages such as HTML and XML as well as on Internet access through ISPs. The second generation of Internet-based services, the Web 2.0, emphasised online collaboration and sharing among users. Social networking sites, wikis, communication tools and folksonomies became the new areas of interest. [11]

The current state of the Internet, Web 3.0, could also be named “intelligent” or “semantic Web”. It comprises semantic technology, Big Data Analysis, machine learning and artificial intelligence. Recent years have witnessed above all a gradual shift of the focus on manually applied semantic technology (manual annotator) to automatically applied processing (automatic annotator) of information. Among others, this includes intelligent applications such as Natural Language Processing, machine reasoning and autonomous virtual agents.

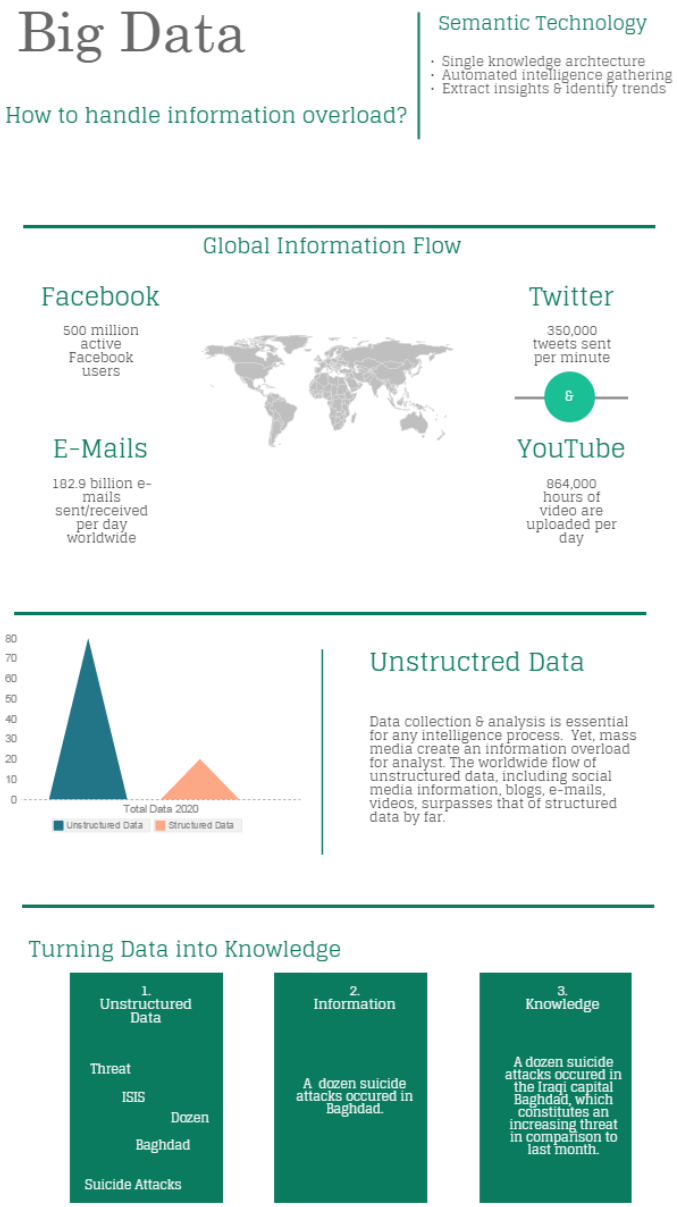
Big Data Analysis

Monitoring the Internet and social media in real time and extracting relevant information to gain applicable intelligence has become one of the biggest challenges for analysts and decision makers over the past years.

Big Data Analysis tools are an integrative element in the decision making process and the day-to-day workflow of large corporate enterprises and international governments. Big Data Analysis pursues one key goal: gathering and quickly generating as well as visualising new information from structured and unstructured data that is centralised in a single corporate knowledge architecture.

The traditional approach to dealing with Big Data (e.g. data mining), however, exhibits clearly visible limits: data is not analysed in its meaning or in its specific context. As a consequence, it is not transformed into new information or even causal knowledge.

The future trend will therefore inevitably be a focus on the fields of semantic technology and artificial intelligence. This approach aims at automatically transforming unstructured data into information and subsequently into applicable knowledge.



Semantic Technology

Software programmes of the Web 2.0 generation recognise and subsequently identify words solely as sequences of letters without understanding the precise context or meaning. The contemporary intelligent Web is capable of crossing this border. It uses semantic technology that is able to understand automatically (1) the structure of words, sentences and texts (letter, word, sentence, paragraph), (2) the causal relationship among words (subject, verb, object) and (3) the contextual meaning of words and/or word combinations.

Example:

1. I like driving my new Jaguar. The 2015 Jaguar is an exhilarating sports car, but our experience with two early models raises concerns about its build.

2. Jaguars hunt mostly at night. The big cat is the third-largest feline after the tiger and the lion, and the largest in the Americas.

3. While driving my jaguar on the paved desert road, I killed a jaguar.

Software that uses semantic technology is able to understand the context of the sentences and accordingly applies the correct meaning of the word jaguar – either as a car or as an animal. Even in the third sentence, where the two words (jaguar) are spelt identically, the system identifies that a car of the brand Jaguar killed a jaguar animal. To this end, semantic software uses language analysis processing. Language analysis enables the system to understand the causal relationship between the two words as well as their meaning and context. More precisely, it differentiates between the Jaguar (as car) and the jaguar (as animal) by understanding that the "car = object = not living" causes the killing of a "jaguar = animal = something that lives and therefore can be killed". By analysing the sentence structure (the sequence of subject, verb, object, etc.), semantic technology can identify "who causes an action" and "who/what is targeted". Semantic systems to a degree can even learn automatically how to use language by employing dictionaries; they are capable of identifying subjects and objects, their actions and relationship to each other, and thus can recognise and depict even complex connections and environments.

Semantic Technology Can Be Used in Manifold Contexts:

In order to highlight the various potentials of applying semantic technology, we have selected some practical examples.

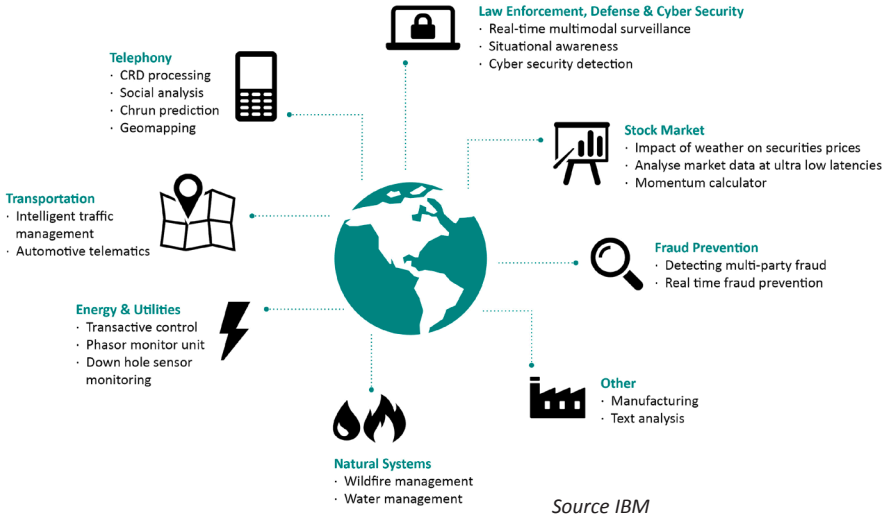
Fraud Investigation: Semantic technology can identify concealed patterns in financial, non-financial and textual data. Patterns include:

- Relationships: Degrees of proximity and/or separation
- Names: Similar and altered names
- Time: Time of transactions (frequency, irregularities)
- Numeric patterns: Correct invoice numbers, fake transaction amounts
- Geographic patterns: Proximity relationships between apparently unrelated entities [12]

Market Analysis: Sentiment/opinion analysis has been used as an integrative part of semantic technology to predict market fluctuation, e.g. via Twitter, Instagram and Facebook. In cases where VIPs communicated about certain products or invested into specific projects/firms, social media distributed the information and caused such a high traffic flow that it resulted in a movement of the market.

Health: Semantic software can support diagnostics by proposing hypotheses about the precise nature of a disease. For example, the technology screens medical libraries and case studies, and compares symptoms or patient stories with other relevant data, thus indicating the most likely diseases.

Advertising: Search machines are increasingly using semantic technology for a more efficient placement of advertisements. The search query of a user looking for a new car will link the potential buyer to advertisements matching his query. With the help of specialised software, his search will point him to the nearest car dealer.



Source IBM

Internet of Things

While large enterprises are still struggling to channel and analyse the growing data flows, more systems are being developed to generate even more data. The new trend of these days is the Internet of Things (IoT). The term mainly describes "a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies" [13]. The IoT creates opportunities for a more direct integration between the physical world and computer-based systems. It allows objects such as cars, Internet-connected smart screen TVs, thermostats, remote power outlets, sprinkler controllers, door locks and home alarms to be sensed and controlled remotely across a network infrastructure.

The consulting firm McKinsey distinguishes among six types.[14]

Information & Analysis			**Automation & Control**		
Tracking Behaviour	**Enhanced Situational Awareness**	**Sensor-Driven Decisions Analytics**	**Process Optimisation**	**Optimised Resource Consumption**	**Complex Autonomous Systems**
Monitoring the behaviour of persons, things or data through space and time	Achieving real-time awareness of physical environment	Assisting human decision making through deep analysis and data visualisation	Automated control of closed (self-contained) systems	Control of consumption to optimise resource use across networks	Automated control in open environments with great uncertainty
Examples: Presence-based advertisements based on locations of consumers	*Example: Sniper detection using direction of sound to locate shooters*	*Example: Continuous monitoring of chronic diseases to help doctors determine best treatments*	*Example: Precise adjustment in manufacturing lines*	*Example: Smart meters and energy grids that match loads and generation capacity in order to lower costs*	*Example: Collision avoidance systems to sense objects and automatically apply brakes*

Semantic technology solutions are an integral part of IoT devices. By now, Natural Language Processing – as a part of semantic technology – includes all main topics, from knowledge representation and reasoning to vision and speech recognition (e.g. face recognition or the transcription of a sound such as our voice). There are a number of new devices on the market that use this technology, such as Google Glass and Microsoft HoloLens. Some "intelligent" devices have already been in use for some time, e.g. Siri on the iPhone. Siri is a virtual assistance programme that interacts with the operator and can understand simple commands, such as:

- "Please arrange an appointment with my dentist on Friday 11 a.m."
- "Please remind me to call my wife at 3 p.m."
- "What is the best route to work" (linkage with car navigation system)
- "Please open the garage door on my return"

Semantic technology is essential in the process of helping Siri understand users' specific search queries. By extracting all grammatical and logical components of the sentence and identifying their function, for example, verbs as a type of action (e.g. search information, or call person, set an appointment) and nouns as objects (appointment, driving route, wife, garage door), programmes like Siri can establish causal relationships between syntactical elements and grasp the meaning of a sentence.

The increasing use of information technology has accelerated the growth of data volumes and will continue to do so in the upcoming years. Natural Language Processing as a part of semantic technology will be an indispensable basis in analysing structured and unstructured data.

Assistance programme Siri

What Can Semantic Platforms Do for Big Data Analysis?

What Can Semantic Platforms Do for Big Data Analysis?

Semantic Big Data Analysis Platforms are used to identify, process and analyse structured and unstructured data, extract causal relationships between words/entities/ concepts and visualise new knowledge. In order to ensure a high degree of performance and accuracy as well as to achieve applicable and useful results, such applications must integrate the following key functionalities.

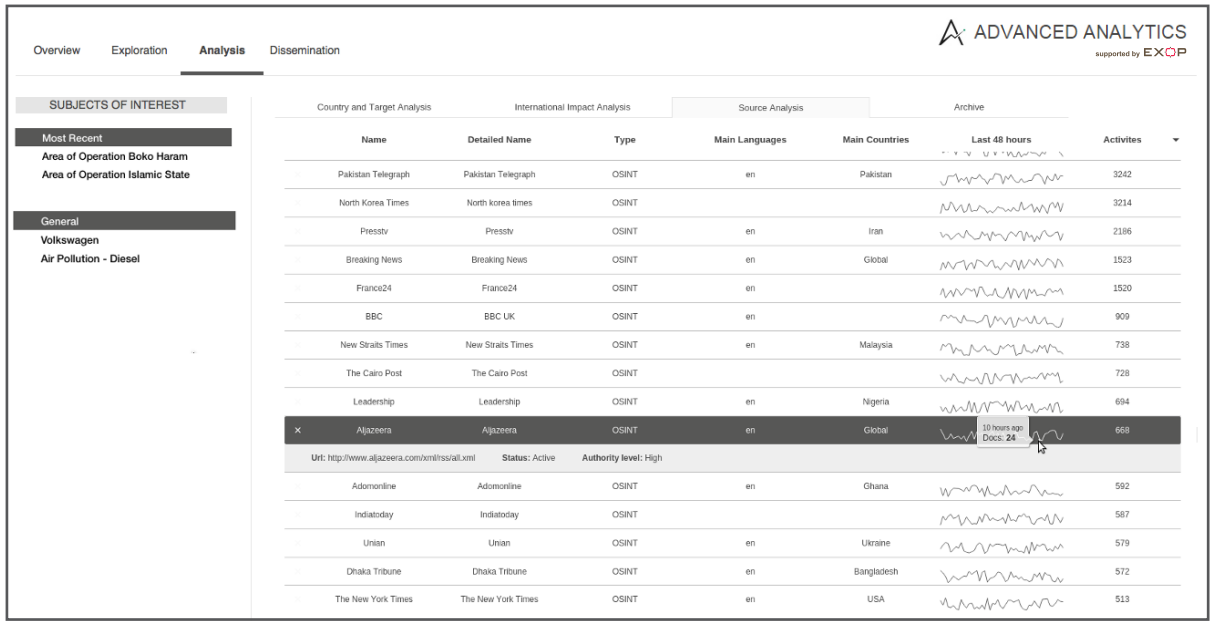
Crawling Structured and Unstructured Data

Relevant information can be collected from several sources, including:

- OSINT: Newspapers, magazines, radio, television and newswires
- SOCMINT: Social networking sites such as Facebook, Twitter, Instagram, etc.; video sharing sites, wikis and blogs
- Official Data: Government reports, official data such as budgets, demographics, hearings, legislative debates, press conferences, speeches, safety warnings and contract awards
- Observation and Reporting: Sightings from amateur aeroplanes, radio monitors and worldwide satellite photography
- Communication: E-mails, phone calls, SMS and applications
- Academia: Academic essays, dissertations and symposia

Most of the information emerges in unstructured form (e.g. digital documents, scanned documents, images and sounds) and is expressed in natural language (e.g. Arabic, English, French, German, Italian, etc.). Semantic software uses text analytics to derive information from these sources. It is used for several purposes, such as:

- Summarisation: Finding the key content within a larger body of information or a single document
- Sentiment/Opinion Analysis: Analysing the attitude (negative, positive, neutral) of a subject (e.g. author, militant/political group, etc.) in regard to a topic or political development
- Smart Search: Obtaining more relevant results by using concepts rather than keywords in search queries
- Classification: Categorising the subject or the key content pieces a specific text deals with



Example: OSINT - basic source analysis, Advanced Analytics/EXOP

Natural Language Processing

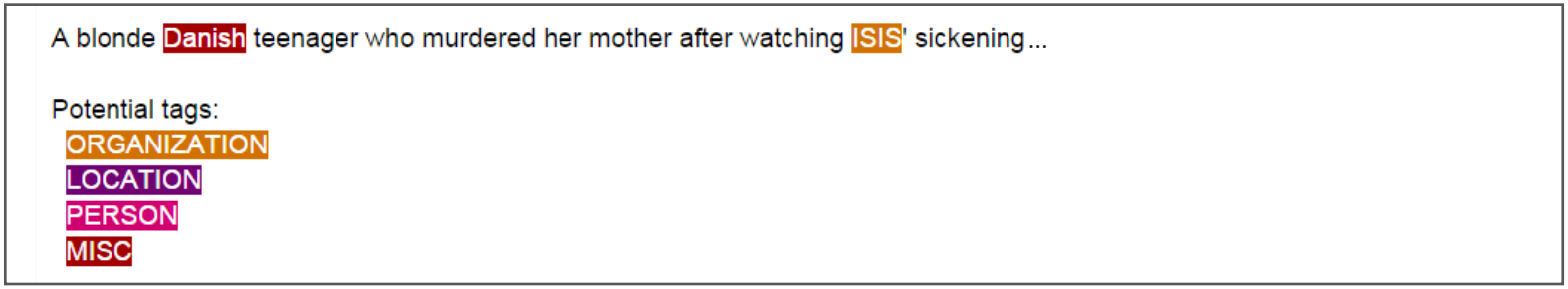
The core of the text analytics process is building Natural Language Processing (NLP) algorithms that use linguistic, statistical and machine learning knowledge to transform unstructured data into relevant information. In general, there are two approaches of NLP: Shallow NLP (the most classical approach) and Deep NLP. Only few platforms use a combination of the two:

1. Shallow NLP

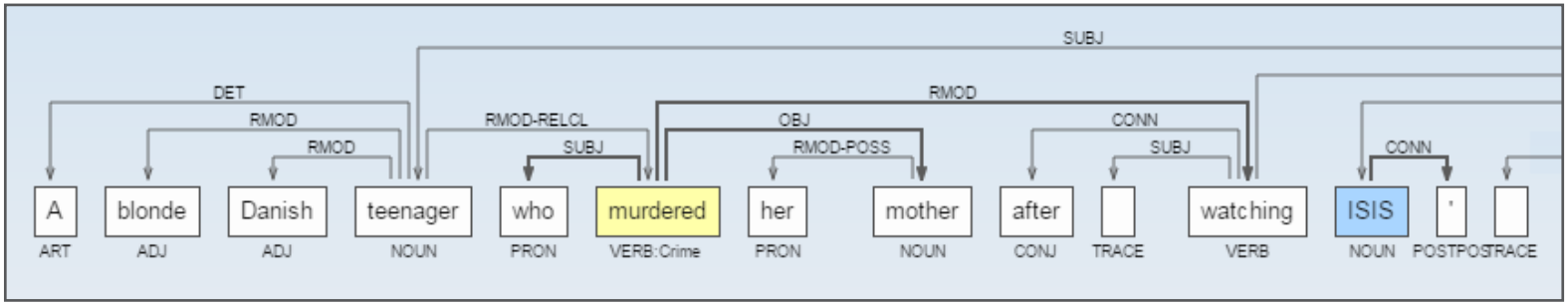
- Needs only training resources (such as tagged corpora)
- Efficiency in terms of calculation
- Extracts the general category of the text
- Generates a partial sentence analysis
- Extracts terms of interest to the domain (person, location, organisation, date)

2. Deep NLP

- Generates a full sentence analysis
- Comprehends the meaning of entire sentences and discourses rather than focusing on text portions alone
- Detects even fine-grained aspects of the language
- Detects implicit information, e.g. entities referred to by expressions such as pronouns



Example 1: Shallow Semantic by Stanford Named Entity Recognizer (NER)



Example 2: Dependency Parser Advanced Analytics/EXOP (Deep Semantic)

Multiple Language Capability

Currently available platforms are able to operate in different languages. The platforms differ from each other in their approach and the accuracy of the expected results of a semantic analysis.

In general, the integration of additional languages is time-consuming, therefore most platforms utilise automatic translation tools (such as Bing or Google translator) in order to develop their own language/semantic capabilities.

This option is appropriate in cases where:

- companies have only very limited time and resources to develop semantic software;
- companies using the translation tools do not necessarily require expertise in the terminology of the subject matter but are satisfied with general, and not complex, wording.

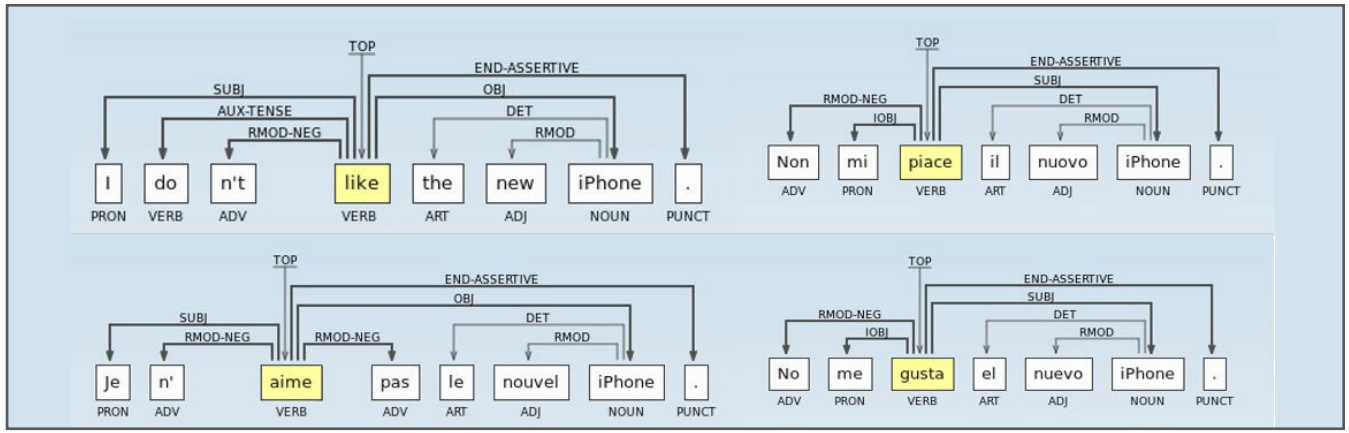
Using standard machine translation provides a rather quick solution but comes with a number of downsides:

- it increases the platform's dependency on third-party translation tools;
- the results of the semantic process lack high-quality accuracy due to error propagation from weak translation performance.

Therefore, some platforms have integrated separate dictionaries and rules for each language instead of relying on machine translation. In traditional approaches, dictionaries and grammatical rules are both constructed manually. Even in a restricted domain (e.g. intelligence), doing this consistently while maintaining a high quality requires non-trivial effort.

Recently, machine learning and in particular Deep Learning technologies have allowed a rapid development of new language modules through "training by reading" of numerous texts. They automatically create a complex parallel list of semantic/conceptual associations in different languages without any additional manual translation effort, e.g. attack (eng) = Anschlag (d) = attacco (it).

The application of Deep NLP leads to more accurate language results. Moreover, by applying grammatical structures and meanings of words/concepts, the user is able to process even the most complex translation request. Most existing platforms, however, only integrate Shallow NLP, which limits the search function basically to extracting names of persons, organisations and locations in different languages.



Example: Multilingual dependency tree, Advanced Analytics/EXOP

Further Functionalities

Event Detection and Clustering

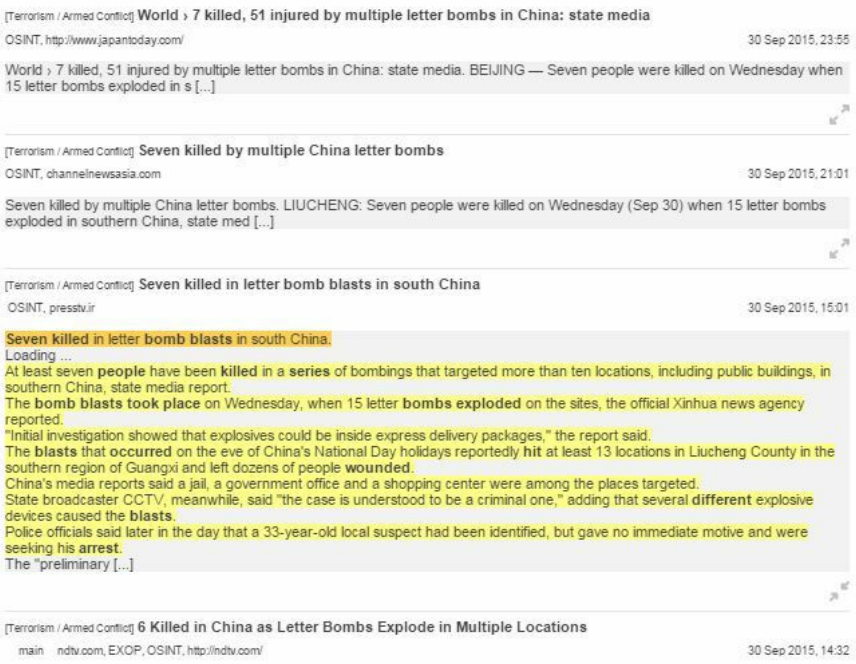
The detection and clustering of single events is a basic requirement for the identification and quantification of risks. Traditional platforms, however, are not able to extract and identify a single incident out of the multitude of reports from different sources. For example, a major bomb attack on a hotel in Mumbai would be widely and intensely noticed and discussed in local, national and international media. Conventional platforms would count and list a large number of newspaper and other reports on the attack. They are not capable of attributing this information to one specific incident, to merge different sources and to integrate and calculate the event as part of the analysis process.

Application Programming Interface

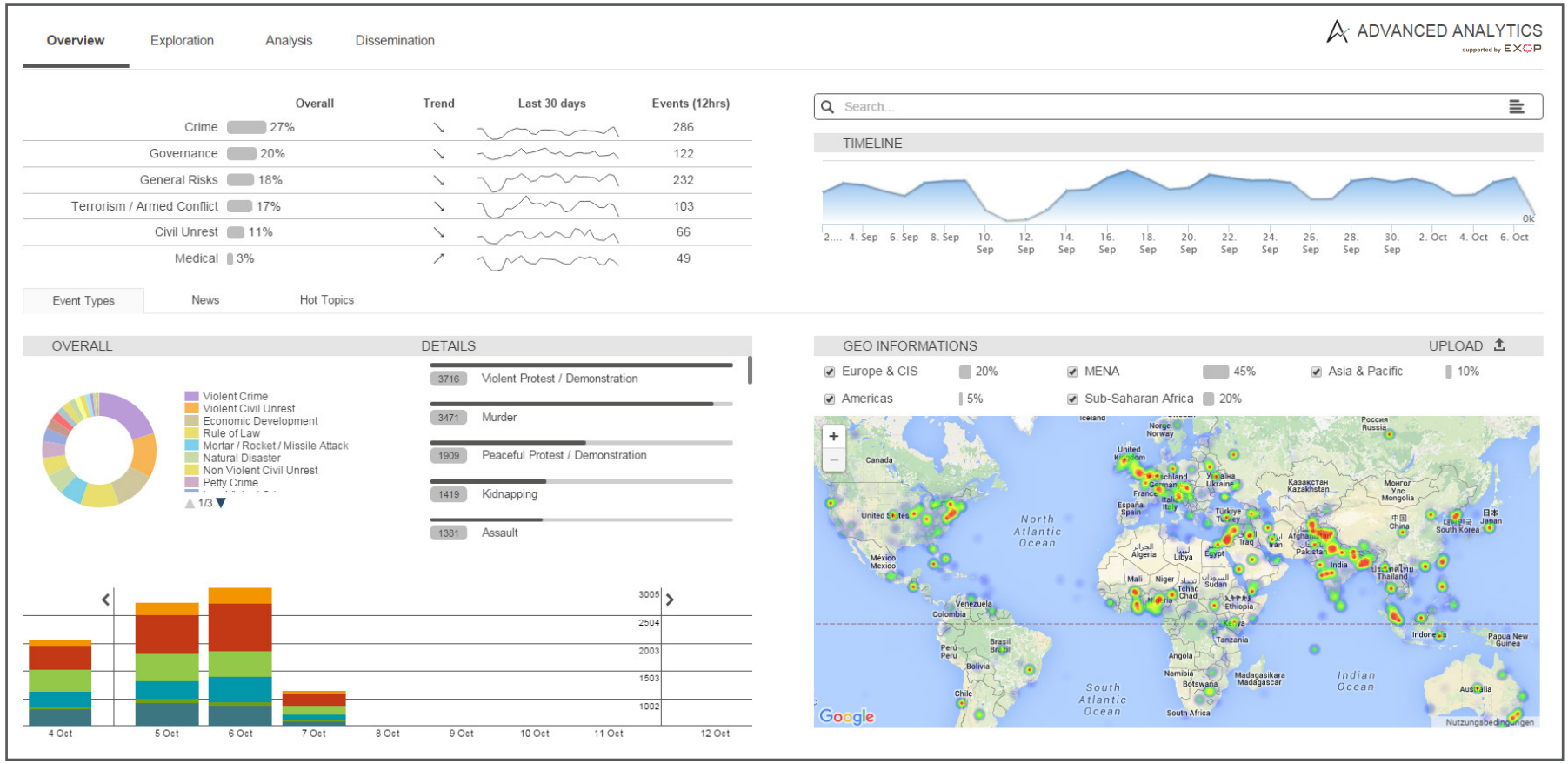
Applying an Application Programming Interface (API) as a service provides the option of exchanging information and/or interacting with third-party software. It allows clients not only to use the graphical interface of a semantic application but also to export data to enrich their projects and applications with Web intelligence.

Automated Reporting

Sophisticated platforms create capacities for analysis because they offer several functionalities with regard to standard reporting. Processing an almost unlimited amount of reports, specific applications are capable of automatically creating e-mails or phone alerts to trigger analysis or operational processes. According to defined Standard Operating Procedures (SOPs), the software can create and export individual reports and analysis products.



Example: Event detection and clustering, Advanced Analytics/EXOP

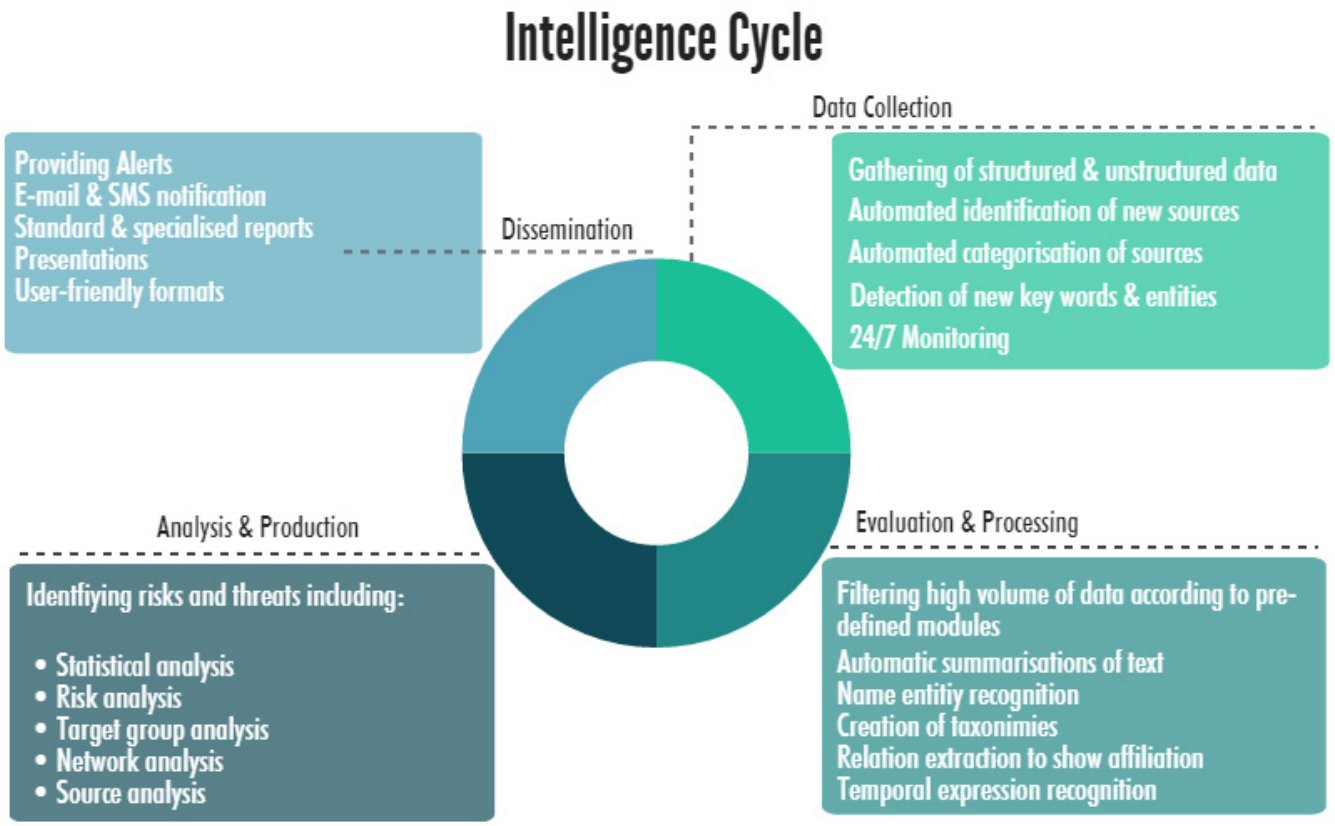


Example: Advanced Analytics/EXOP Intelligence Platform

Benefits of Semantic Intelligence Platforms to the Intelligence Process

Benefits of Semantic Intelligence Platforms to the Intelligence Process

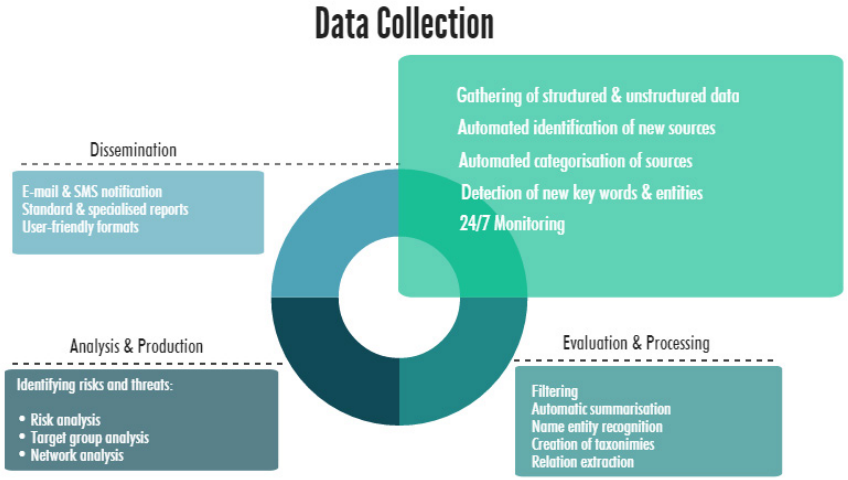
Whether anti-fraud investigations, compliance enquiries or analysing financial, security or general risks, risk and threat assessments are conducted in all cases and follow the same procedure: (1) analysts gather data, (2) analysts evaluate, structure and process information, (3) subject matter experts analyse causal relationships and identify relevant risks. Finally, at the end of this process, (4) an analysis is issued. Semantic technology can assist analysts in every step of the intelligence cycle by increasing performance and improving accuracy.



How Semantic Technology Benefits Data Collection

Task:

Data collection is the initial stage of any intelligence process. This includes structured data such as statistical reports and country reports as well as unstructured data in different formats, e.g. social media information, e-mails, images and videos. Over the next years, unstructured data will constitute the majority of the available data corpus and most likely become the most valuable source of information. Processing and controlling the real-time inflow of structured and unstructured data from various external and internal sources poses a major challenge for every form of analysis.



Key Challenges for Analysts:

- Quickly finding and collecting structured and unstructured relevant data, including the monitoring of broadcasts, audio files, newswires, social media, RSS feeds, etc.
- Integrating all relevant external and internal data in different formats into a single database for subsequent evaluation, structuring and analysis
- Time-efficient source management

Benefits of Semantic Technology:

1. Semantic technology can centralise data by creating a single data architecture.

Sophisticated platforms operate as an umbrella for external and internal data, libraries, e-mail archives, analysis and fact finding databases. They also offer an interface for integrating data of external providers as well as interconnecting with third-party analysis tools.

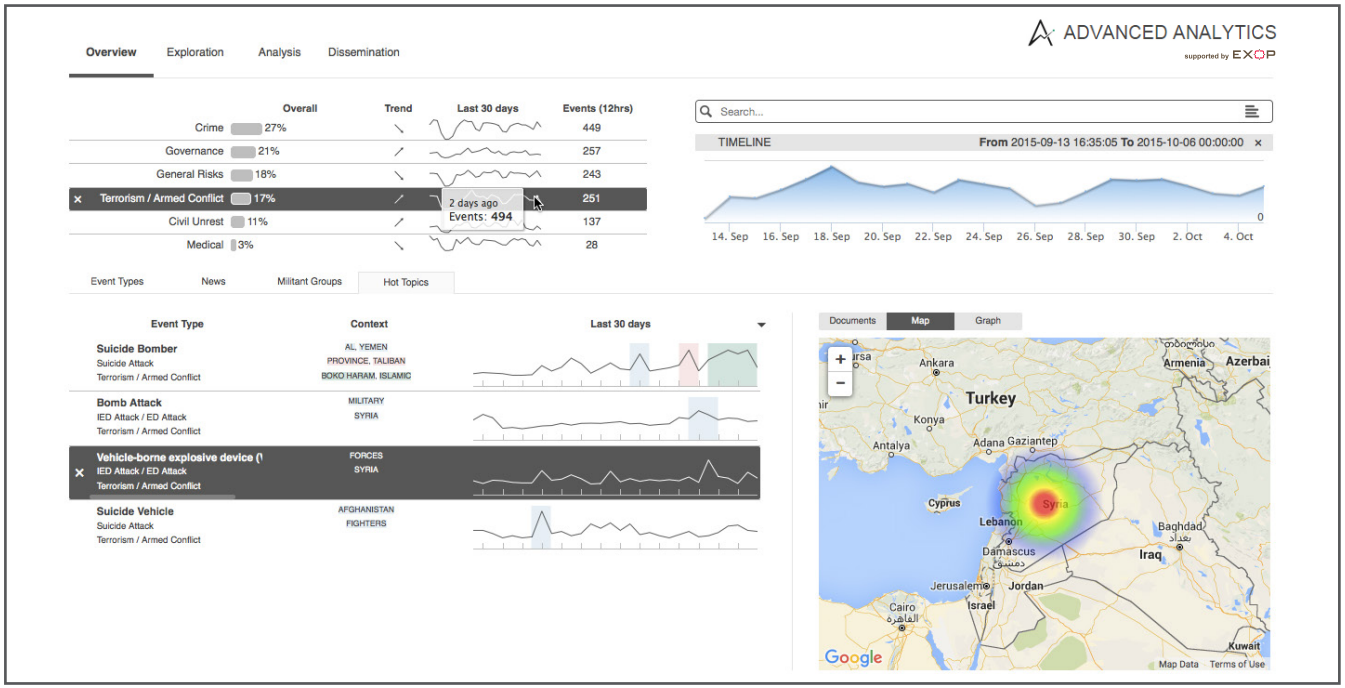
2. Semantic technology platforms support experts in the identification, selection and management of sources.

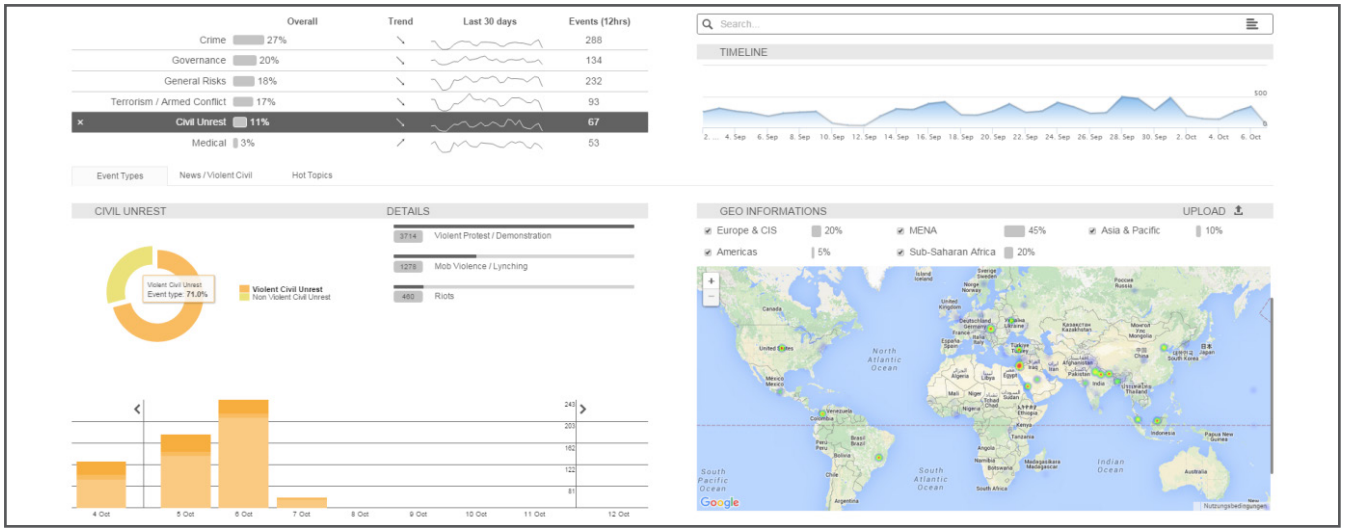
Social media accounts are created, relocated or deleted at a rapid pace. Semantic software identifies changing accounts and suggests new relevant sources (e.g. Twitter accounts, RSS feeds, blogs) according to the interest of the user. A semantic platform also prioritises and categorises sources according to the accuracy and reliability of data as well as the distinctive region, language and/or topic of interest.

3. Semantic technology automatically suggests new relevant keywords, concepts, groups and classifications.

By systematically crawling data 24/7, semantic systems are able to recognise the emergence of new relevant keywords and concepts that could be of specific interest to the user. Semantic technology platforms can automatically detect:

- new modi operandi used by militant groups (single suicide operations replacing remote-controlled car bombs);
- an emergence of a new militant group out of a mother organisation;
- new behavioural patterns of monitored subjects and entities;
- a changing area of operations.



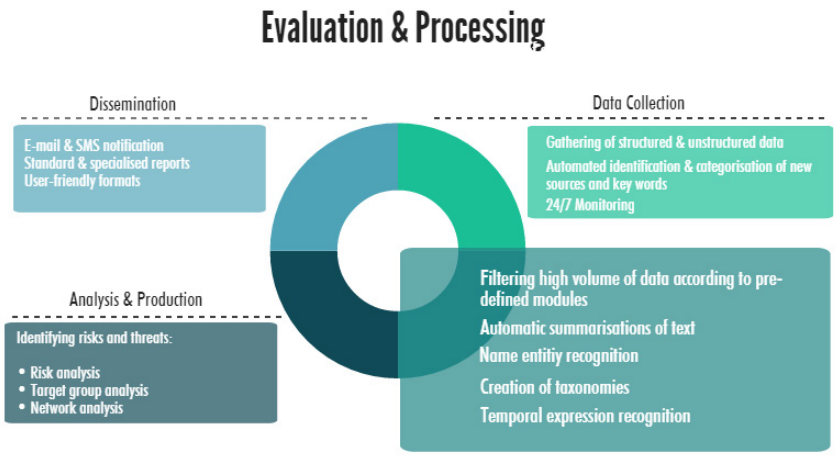


Example: Advanced Analytics/EXOP Intelligence Platform

How Semantic Technology Benefits Evaluation & Processing

Task:

After the data has been collected and the information extracted, the key details are processed for exploitation. This involves the translation of raw intelligence material from a foreign language and the evaluation of its relevance and reliability. Raw intelligence is collated in preparation for exploitation.



Key Challenges for Analysts:

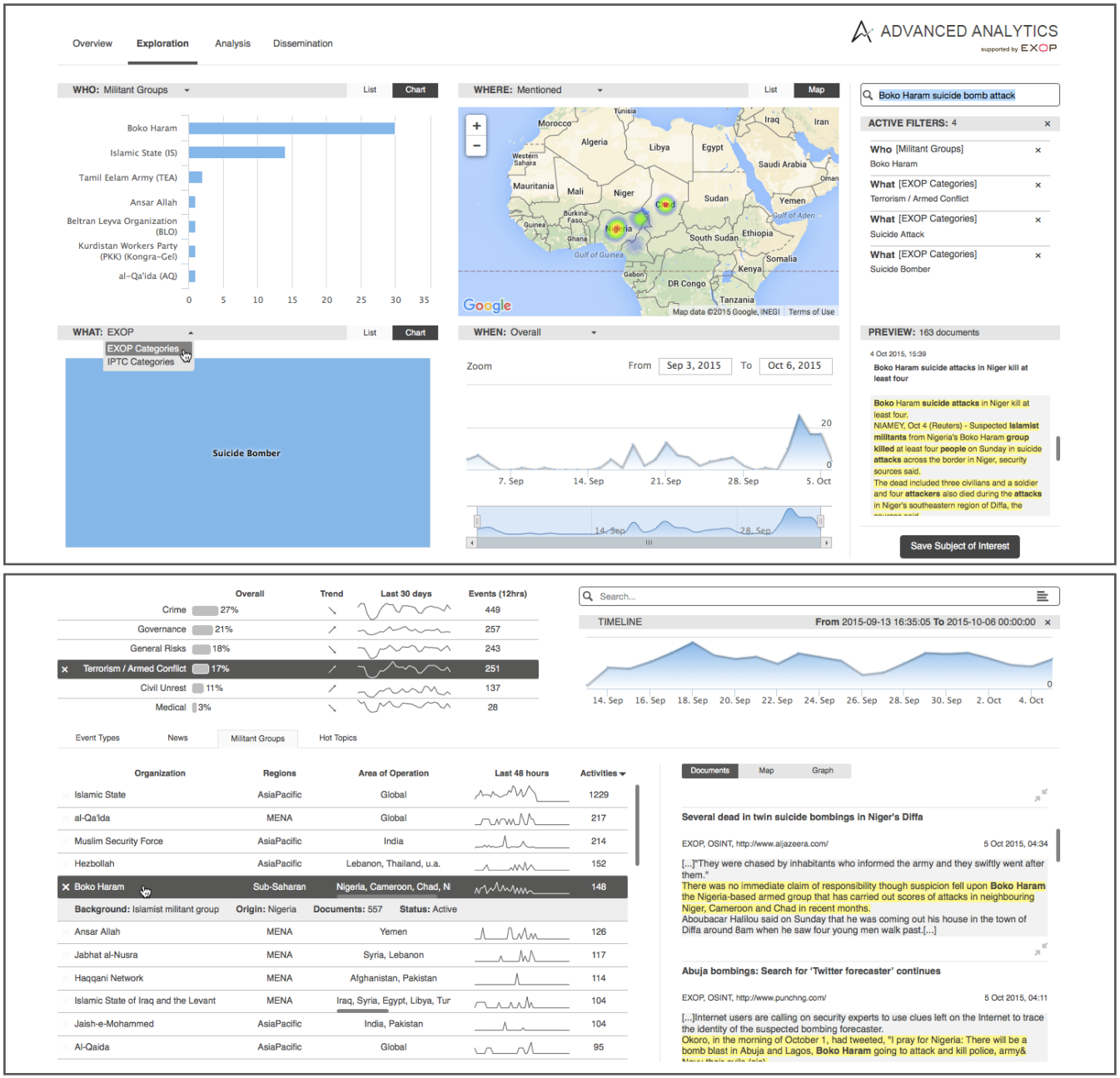
- Prioritising and categorising collected data quickly, and processing the gained relevant information for further in-depth analysis

Benefits of Semantic Technology:

Semantic technology allows the filtering of high data volumes according to predefined modules, e.g. languages, subjects of interest, regions of interest and specific time frames. Filtering the data provides results of high accuracy and enables the subsequent prioritisation of information, thus isolating the most relevant and critical information.

Semantic technology automatically applies Natural Language Processing (NLP), which facilitates the evaluation process by means of:

- automatic summarisations of texts in order to make relevant information visible;
- machine translations (Deep/Shallow NLP) for the comprehension of complex sentences and subject matter terminology in different languages (for more details, see page 23);
- name entity recognition to locate and classify text elements in predefined categories, such as the names of persons, organisations, locations, quantities, monetary values, percentages, etc.;
- relation extraction to show affiliations, links and relationships between monitored subjects or entities of interest;
- opinion detection to identify and analyse changing perceptions and interpretations of political decisions or other developments;
- temporal expression recognition to analyse persons, things or any data through space and time.



Example: Advanced Analytics/EXOP Intelligence Platform

How Semantic Technology Benefits Analysis & Production

Task:

Analysts assess the significance and implications of processed intelligence. They integrate this data by combining disparate pieces of information to identify collateral information and patterns. Finally, the significance of any newly developed knowledge is assessed.

Key Challenges for Analysts:

- Developing available information into knowledge by identifying risks, threats, patterns and trends

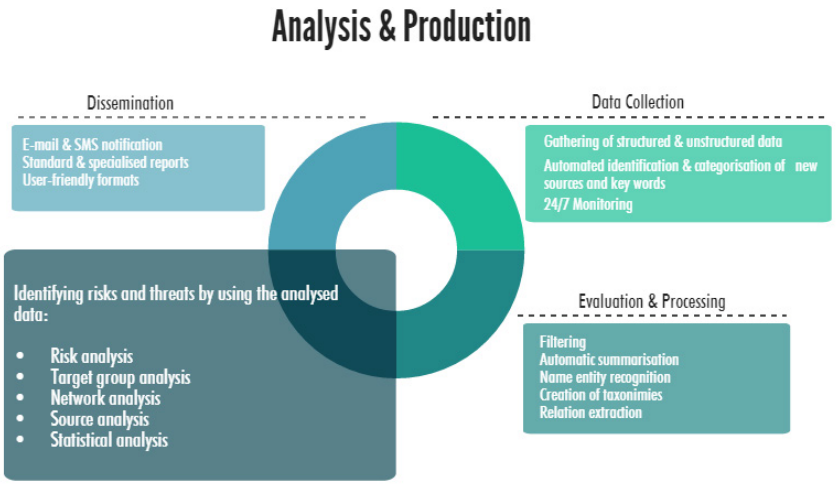
Benefits of Semantic Technology:

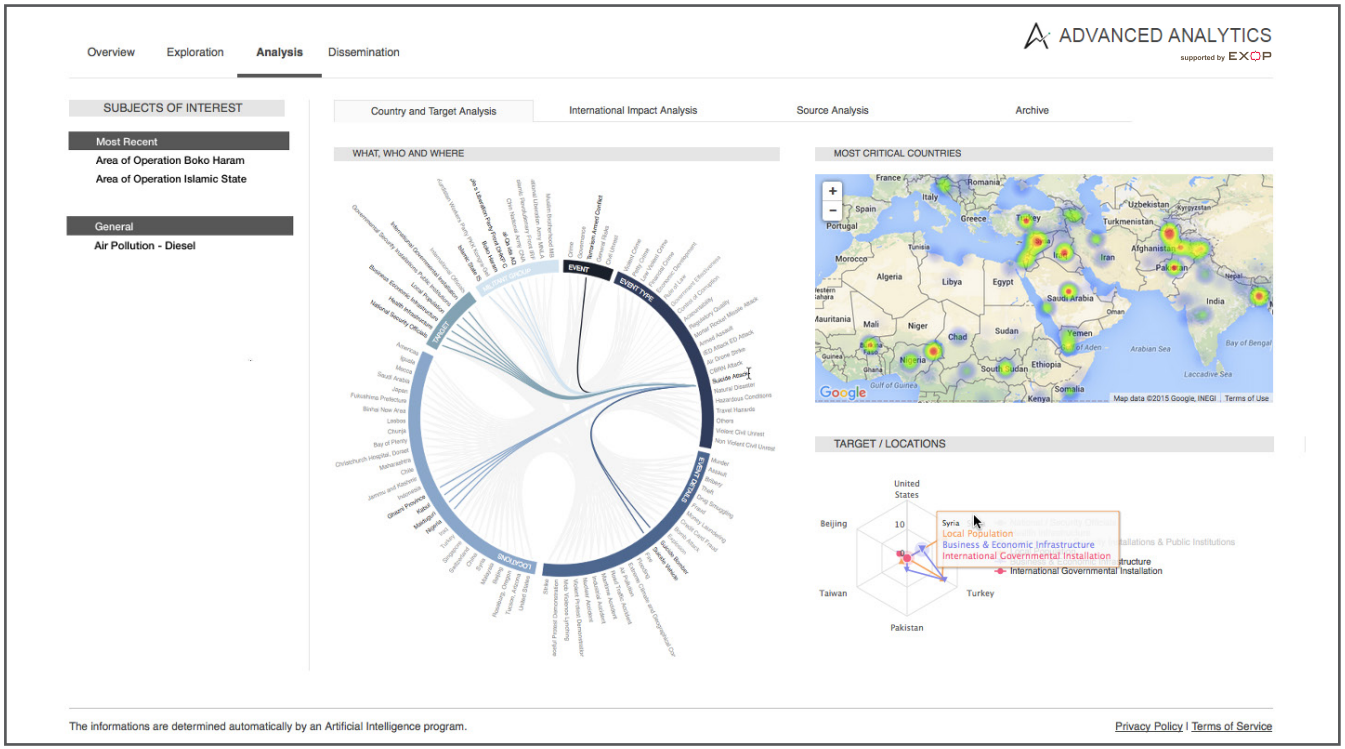
Semantic technology platforms constitute a single knowledge architecture. They extract information from all external and internal sources, including qualitative and quantitative data in different formats (html, PDF, xml, word, excel), that have been integrated into the system.

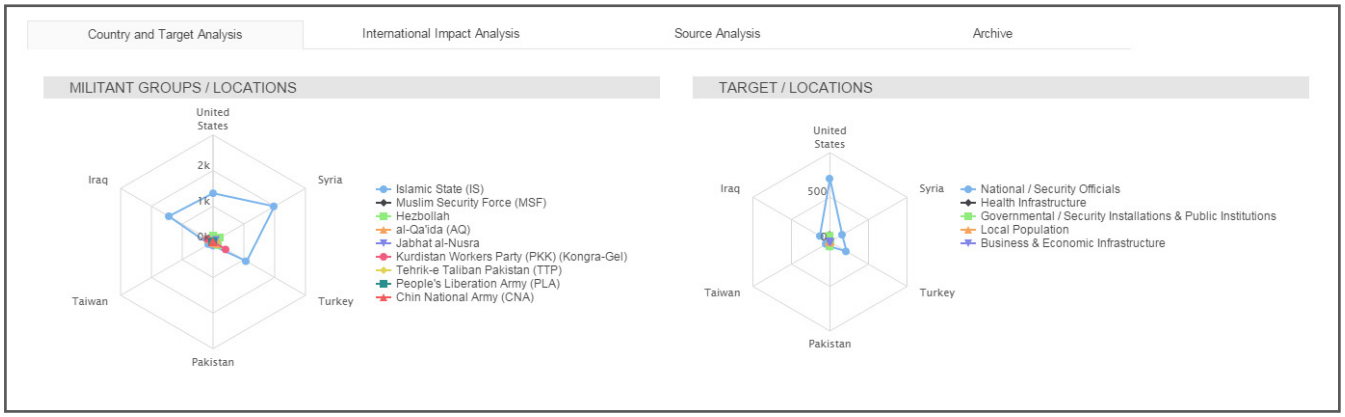
While exploring the available information, these applications automatically generate new insights by visualising relevant data.

Semantic software automatically identifies trends by using the analysed data. It visualises identified patterns in maps and configurable charts and finds application in:

- Statistical analysis
- Risk analysis
- Target group analysis
- Network analysis
- Source analysis







Example: Advanced Analytics/EXOP Intelligence Platform

How Intelligent Dissemination Can Further Enhance the Benefits of Semantic Technology

Task:

Sending pertinent results and reports to relevant authorities as a basis for decision making – while this is not necessarily a primary task of semantic technology, a semantic platform can still support the dissemination process.

Key Challenges for Analysts:

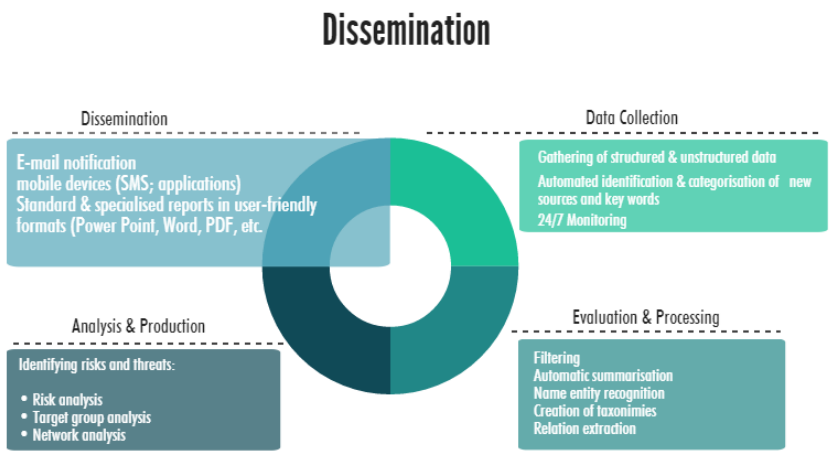
- Aligning reporting requirements to the level of urgency
- Sending information via preset dialogue channels to a list of recipients in specific time intervals (e.g. 6.30 a.m. and 8.30 a.m. 24/7/365)
- Making reports available in different formats (mobile devices, desktop version) and for dominant operating systems (Blackberry, Android, Microsoft)
- Tracing key sources/quotes/full articles that serve as a basis for reports

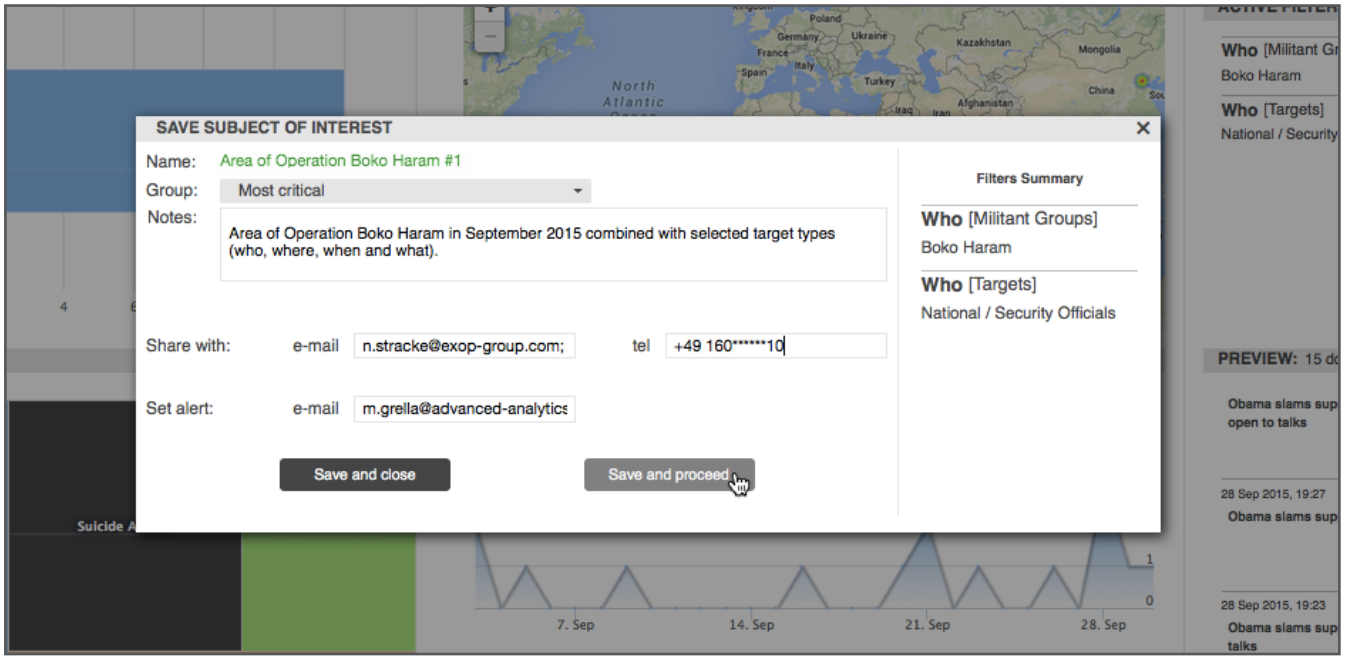
Benefits of Semantic Technology:

Sophisticated platforms issue alerts and standard reporting via:

- e-mail notification
- mobile devices (e-mail, SMS, applications)
- user-friendly formats for reports (Power Point, Word, PDF, etc.)

Individualised rights management systems (admin/user levels) allow for an automated and manual issuance of specific reports while keeping in consideration the distinctive security classifications of recipients/users. Semantic platforms can be adapted to external operating systems with minor programming efforts.





Example: Advanced Analytics/EXOP Intelligence Platform

Key Requirements on Technology and Information Security

Key Requirements on Technology and Information Security

Semantic platforms can be accessed via web portals hosted on a secure website or be installed in-house. There are a number of common requirements that apply to the basic functionalities, technology and infrastructure of semantic applications.

Intelligence Process	*Functionality*	*Technology*	*Information Security*
Collection	Automatic identification of new sources Automatic categorisation of sources and related documents Detection of new keywords and entities Provides single-source architecture for OSINT, SOCMINT and heterogeneous sources	Harvests information with multi-source data retriever Architecture: sources (RSS feeds, crawler, grey literature, special source) > indexed documents Single-incident identification through fact, meaning and document similarity: documents > event Improves processing by using NLP, including: • IPTC and OSINT-based categorisation • Specific taxonomy of terms and concepts for intelligence (including Armed Conflict, Terrorism, Crime)	Secure by Design • All access to the Web platform should be secured via TLS • Search queries should only be saved on explicit user request, otherwise queries are not recorded • External content should only be indexed for the purpose of a full text search; no copy of potential harmful sources is stored on the platform; additionally, all content passes through sanitiser module, minimising the risk of infection IT Infrastructure • Platform operator should employ an Information Security Management System (ISMS) such as ISO 27001 • Implement the security guidelines outlined in Federal Office of Information Security (BSI) „IT-Grundschutz Manual“ • Ensure scalability and high availability by distributing critical system components across multiple network nodes • System should be hosted in country with strong privacy protection laws, such as Germany • Payment information and credit card details should be stored on a separate server in an encrypted database Software Development Lifecycle • Threat modelling and mitigation/attack surface reduction during design • Developer training/awareness for secure coding practices • Code reviews/testing/static analysis/fuzzing • Penetration testing • Use of strong cryptography to protect sensitive information (e.g. passwords) • Multi-tenancy capabilities to separate customer data
Evaluation & Processing	Filters & prioritises high volume of data according to individual needs Improves processing by using NLP	Identification and ranking of relevant information according to arbitrary subject of monitoring through semantic filter combination: • “Google-like” search • Faceting search • Auto-detecting relevant phrases Improving processing by using NLP, including: • Automatic summarisations of text • Name entity recognition • Relation extraction • Opinion detection • Temporal expression recognition	
Analysis & Production	Develops information as a single knowledge architecture by: • Extracting new insights from information • Analysing forecasts • Identifying trends • Integrated analysis functionalities • Integration of internal and external sources with qualitative & quantitative data	Geo-spatial events decoding Automatic impact/risk evaluation Predictive algorithms Application of inference engine using “Analyst Prior Knowledge”	
Dissemination	Issuance of alerts & standard reporting according to Standard Operating Procedures	Data exchange library for export insight into software/infrastructure of third parties Advanced engine that automatically detects alerts and warnings Data export to MySQL, elastic search (or fast document-based database), QlickView or similar BI tools and other RDMS	

Annex

Annex

Footnotes

[1] See *Understanding Terrorism Innovation and Learning.*

[2] For a more comprehensive analysis of this topic, see Atwan, *Islamic State*.

[3] Veilleux-Lepage, *Retweeting the Caliphate*.

[4] See *Mexico's Struggle for Public Security* and *Transnational Organized Crime*.

[5] See Vázquez, *México, primer lugar en penetración Social Media*.

[6] See Randal, *Mexican Helicopter Shot Down* and Córdoba, *Mexican Army Helicopter Was Shot Down*.

[7] See *From Venus to Mars*?

[8] For an unclassified overview, see *Intelligence Preparation of the Battlespace*.

[9] See *Local Politics in Afghanistan*.

[10] See the critical assessment of Münch, *Wahrnehmung und Analyse der Taliban.*

[11] http://blog.thedigitalgroup.com/sheldonf/2015/04/23/introduction-to-web-3-0-the-semantic-web/.

[12] https://chapters.theiia.org/topeka/ChapterDocuments/Fraud%20and%20Advanced%20Data%20Mining.pdf.

[13] See *Overview of the Internet of Things.*

[14] http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_things.

Bibliography

- Atwan, Abdel Bari: Islamic State. The Digital Caliphate. Oakland, CL: University of California Press, 2015
- Berger,J.M., and Jonathon Morgan: The ISIS Twitter Census. Defining and describing the population of ISIS supporters on Twitter. The Brookings Project on U.S. Relations with the Islamic World. Analysis Paper 20, March 2015
- Córdoba, José de: Mexican Army Helicopter Was Shot Down With Rocket-Propelled Grenades. The Wall Street Journal, 4 May 2015, http://www.wsj.com/articles/mexicanarmy-helicopter-was-shot-down-with-rocketpropelled-grenades-1430783784
- Detailed Analysis of Islamic State Propaganda Video: Although the Disbelievers Dislike It. Quilliam Foundation Presentation; http://www.quilliamfoundation.org/wp/wp-content/uploads/publications/free/detailed-analysis-of-islamic-state-propaganda-video.pdf
- From Venus to Mars? Provincial Reconstruction Teams and the European Military Experience in Afghanistan, 2001–2014. Ed. Bernhard Chiari in collaboration with Thijs Brocades Zaalberg, Nicola Labanca and Ben Schoenmaker. Freiburg i.Br.: Rombach, 2014
- Grella, Matteo: Notes About a More Aware Dependency Parser, arXiv, 20 July 2015, http://arxiv.org/pdf/1507.05630v1.pdf
- Intelligence Preparation of the Battlespace. Ed. (US) National Geospatial-Intelligence College, http://www.dsg.eb.mil.br/seminario_eua/arquivos/H-010%20Intelligence%20Preparation%20of%20the%20Battlespace.pdf.
- Internet of Things. Mc Kinsey Quarterly, April 2010, http://www.mckinsey.com/insights/high_tech_telecoms_internet/the_internet_of_thing
- Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems. Ed. Ovidiu Vermesan and Peter Friess. Aalborg: River Publishers, 2013
- Islas, Octavio: Cifras sobre jóvenes y redes sociales en México. Entretextos 7 (19), April - July 2015, http://entretextos.leon.uia.mx/num/19/PDF/ENT19-1.pdf
- Jurafsky, D., and J. H. Martin: Speech and Language Processing. An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics. 2nd edition. Upper Saddle River, NJ: Prentice-Hall, 2009
- Local Politics in Afghanistan: A Century of Intervention in the Social Order. Ed. Conrad Schetter. London: C Hurst & Co Publishers Ltd, 2013
- Manning, Christopher D., and Hinrich Schütze: Foundations of Statistical Natural Language Processing. Cambridge/Mass: MIT press, 1999
- Mexico's Struggle for Public Security. Organized Crime and State Responses. Ed. George Philip and Susana Berruecos. New York: Palgrave Macmillan, 2012
- Münch, Philipp, »Wahrnehmung und Analyse der Taliban durch die Bundeswehr«, in: Der Taliban-Komplex. Zwischen Aufstandsbewegung und Militäreinsatz. Ed. Conrad Schetter and Jörgen Klußmann. Frankfurt a.M.: Campus, 2011, pp. 229 - 244
- NATO's Open Source Intelligence Handbook (2001), http://www.oss.net/dynamaster/file_archive/030201/ca5fb66734f540fbb4f8f6ef759b258c/NATO%20OSINT%20Handbook%20v1.2%20-%20Jan%202002.pdf
- Noubours, Sandra, Albert Pritzkau, and Ulrich Schade: NLP as an Essential Ingredient of Effective OSINT Frameworks. Conference Paper. Fraunhofer Institute for Communication, Information Processing and Ergonomics (FKIE), 2013, http://publica.fraunhofer.de/dokumente/N-282073.html
- Overview of the Internet of things. Ed. International Telecommunication Union, Series Y: Global Information Infrastructure, Internet Protocol Aspects and Next Generation Networks. Next Generation Networks – Frameworks and functional architecture models. Y.2060 (06/2012), https://www.itu.int/rec/T-REC-Y/en
- Randal, C. Archibold: Mexican Helicopter Shot Down, Killing 3 Soldiers. The New York Times, 01 May 2015, http://www.nytimes.com/2015/05/02/world/americas/mexicanhelicopter-shot-down-killing-3-soldiers.html
- Russell, Stuart J., Peter Norvig: Artificial Intelligence: A Modern Approach (3rd edition). Upper Saddle River: Prentice Hall, 2010
- Simcox, Robin: AQAP's Ideological Battles at Home and Abroad. Current Trends in Islamist Ideology 18, May 2015, pp. 18-40
- Transnational Organized Crime. Analyses of a Global Challenge to Democracy. Ed. Heinrich-Böll-Stiftung and Regine Schönenberg. Bielefeld: transcript Verlag, 2013
- Understanding Terrorism Innovation and Learning: Al-Qaeda and Beyond. Ed. Magnus Ranstorp and Magnus Normark. New York: Routledge, 2015
- Woody-Coussens, Shauna, and Jeremy Clopton: Secrets, Conspiracies and Hidden Patterns. Fraud and Advanced Data Mining. BKD Presentation, https://chapters.theiia.org/topeka/Chapter-Documents/Fraud%20and%20Advanced%20Data%20Mining.pdf
- Vázquez, Rubén: México, primer lugar en penetración Social Media. Forbes Mexico, 31 July 2014, http://www.forbes.com.mx/mexico-primer-lugar-en-penetracion-social-media/
- Veilleux-Lepage, Yannick: Retweeting the Caliphate: The Role of Soft-Sympathizers in the Islamic State's Social Media Strategy. Conference Paper Presented at the 6th International Symposium on Terrorism and Transnational Crime. Antalya, Turkey. December 4-7, 2014

Internet addresses and URLs were rechecked for the last time 5 October 2015.

Abbreviations

AI	Artificial Intelligence
AQAP	Al Qaeda in the Arabian Peninsula
IoT	Internet of Things
IPB	Intelligence Preparation of the Battlespace
ISAF	Security Assistance Force
ISP	Internet Service Provider
NATO	North Atlantic Treaty Organization
OSINT	Open Source Intelligence
SOCMINT	Social Media Intelligence
IS	Islamic State of Iraq
NLP	Natural Language Processing

Contact

AA | Advanced Analytics GmbH is a risk management consultancy specialising in the development of semantic intelligence platforms. The combination of technical and analytical expertise provides a profound knowledge base for the creation of intelligence platforms that can identify, analyse and forecast prevalent risks and emerging threats.

We develop tools that assist in finding solutions for the complex problems that are a challenge for analysts and decision makers in their everyday work. Our tools pursue one goal: to simplify the work and the procedures in the analysis process. Collecting real-time data, visualising high-priority information and creating new knowledge are the key components of our semantic technology.

EXOP | The EXOP GmbH is German risk management consultancy that enables clients to establish or further develop a comprehensive risk management for business activities abroad.

Preventive trainings, 24/7 assistance service, worldwide situation reporting for travellers and country risk analysis are trusted services. Professional crisis management and profound expertise in the field of research and analysis belong to EXOP's core competences.

Advanced Analytics GmbH

Byk-Gulden-Straße 24
78467 Constance
Germany

www.advanced-analytics.eu

info@advanced-analytics.eu

EXOP GmbH

Byk-Gulden-Straße 24
78467 Constance
Germany

www.exop-group.com

info@exop-group.com