Auteurs: Saman Tamo en Miguel Chehin. Saman Tamo is sinds 2015 docent-onderzoeker cybersecurity aan de Haagse Hogeschool. Hij is sinds 2022 ook actief als ondernemer bij CTRL Disrupt, waar hij met zijn collega's een bijdrage wil leveren aan een digitaal veilige samenleving en kennisdeling op het gebied van cybersecurity. Hij is te bereiken via: Saman@ctrl-disrupt.nl. Miguel Chehin is werkzaam als security consultant bij CTRL Disrupt. Hij is te bereiken via: Miguel.chehin@ctrl-disrupt.nl.

Context Driven Data Gathering Framework

Data Gathering is an often overlooked step in the process of improving data analysis. Data analysts go through countless data streams whilst trying to filter relevant from irrelevant data. This data analysis process is time consuming and costly, often resulting in analysts spending the majority of their time on filtering out irrelevant data instead of focusing on analysing the data that is important to their organisation. In this framework we aim to provide structure to the process that foregoes data analysis, namely *data gathering*. To improve the quality and relevancy of the data you collect and thereby minimize the strain on data analysts.

When analysts are exposed to a huge amount of irrelevant material, analyst fatigue can develop. This condition may lead to fatigue, burnout, and decreased productivity, which could have detrimental effects on businesses (1). Organisations may suffer major repercussions because of analyst fatigue brought on by irrelevant data. It may result in lost opportunities, incorrect data interpretation, and subpar decision-making. Thus, it is crucial that businesses take action to reduce this issue (2).

There are several areas where you can find sources of unrelated data. Open-source intelligence (OSINT) feeds, for instance, might contain a lot of noise and misleading information when used for intelligence analysis. Users may upload inaccurate or irrelevant information on social media platforms, which can be another source of irrelevant data. Furthermore, a study from Oxford University shows that false positive triggers in event management systems such as SOCs and SIEMs could be responsible to false positive rates of up to 99%, thereby severly increasing the incoming amount of raw irrelevant data (3).

Whilst the focus on reduction in false positive rates and filtering of (ir)relevance in data has primarily been done on the data analysis side, using methods such as, but not limited by:

- data clustering
- pattern tracking
- regressions
- predictions

We find that organisations should spend more effort on the step that priors' data analysis, therefore trying to up the quality and relevance of the data before any data analysis is applied.

Mapping your organisational context involves applying three levels of situational awareness in dynamic environments where decisions must be made frequently.

It should be noted that Context Driven Data Gathering does not eliminate or replace the need of further data analysis and filtering, but should be used to complement the analysis part and assist the analyst in working more efficient and effective.

By establishing and defining the context of your organisation, a well-designed method for gathering data can help your organisation to reduce irrelevant data. Establishing specific goals will help to focus data collection efforts on the most important information and lessen over-saturating analysts with unnecessary data (4).

This strategy should prioritize data sources based on your organisational needs, that are most likely to have pertinent information. Thus, this prioritization is likely to reduce the volume of irrelevant data that analysts must filter through, which would lead to reduced time spend on analysing what is not relevant to your organisational needs (5).
In order to define data as pertinent or irrelevant it is important to lay the grounds on what data could be seen as important in the first place. This framework will focus on assisting you in mapping the organisational context and possible relevance of data based on the importance it has for the enterprise levels within your organisation. This Framework distinguishes 3 enterprise levels into Strategic, Tactical and **O**perational (**STO**).

Step1 Requirements

The requirements stage is critical to a successful Context Driven Data Gathering process. During this phase, the team determines the intelligence program's objectives and operating procedures based on stakeholder requirements. Organisational context and situational awareness are essential to proper data gathering. Mapping your organisational context involves applying three levels of situational awareness in dynamic environments where decisions must be made frequently. A proposed way to map your organisational context is to apply three levels of **S**ituational **A**wareness (**SA**) (6)(7).

Level 1 SA: Perception of the Elements in the Environment

It is important to be aware of what is needed and present within your organisation to perform (daily) tasks and determine what key infrastructures one is dependent on.

To map the elements within your organisation you can start by logging:

- what kind of software is our organisation using, such as applications and operating systems;
- what types of hardware is the organisation using;
- which of those elements are locally managed versus outsourced and
- what could be the possible attack surface of your organisation.

The Importance of the perception of these environmental elements became clear when the renowned log4shell zero day was used to execute arbitrary code in Apache's log4j, with countless organisations not being aware of the fact that they were making use of Apache's log4j in the first place, thus being vulnerable to the exploit (8)(9).

Level 2 SA: Comprehension of the Current Situation

After your organisational perception of elements are consciously mapped and documented, it becomes of importance to comprehend and understand your organisational needs. This way, you are able to change the perimeters of data you are willing to find and focus on what is relevant and important to your organisation.
One way to do that, is to break the organisation down into 3 enterprise layers (**STO**): **S**trategic **- T**actical – **O**perational. The concept behind this approach is to acquire data more consciously, it is crucial to be aware of the needs and desires of your company at several levels (strategic, operational, and tactical).

Efficiency: you can concentrate on gathering data that is pertinent and practical when you have a clear understanding of the data requirements for your firm. By preventing the collecting of unnecessary data, you can save time and money. Accuracy: being aware of the precise information that is wanted at various levels, helps you gather the proper data to satisfy those needs. This implies that the data you gather will be more relevant and beneficial for making decisions and conduct analyses on. Relevance: as each level of the organisation has different information needs, it is crucial to gather information that is pertinent to that level.

Level 3 SA: Projection of Future Status

Level 3 situational awareness involves predicting how the environment will behave and impact your organisation in the near future. Understanding the dynamics and status of the elements at Levels 1 and 2 is necessary for this. For example, if your organisation uses a certain Linux Distribution, you can search for vulnerabilities or news relevant to it. Also, if the strategic layer of your organisation plans to do business with foreign countries, you should monitor for possible conflicts that could affect the mission and vision of your organisation. Without situational awareness, crucial data could be missed.

Step 2 Data Collection

Finding the right data collection tools based on your organisational context is important, because it allows an organisation to tailor its data collection process to its specific needs and goals. Different tools are used to collect different types of data from different sources, while the list of possible data collection methods and tools is very large, we would

like to give you a few tooling suggestions to cater your data gathering needs. It is important to understand that tools can be used to gather data on all enterprise levels and are not limited to a single **STO** enterprise level.

We recommend looking at the GitHub Repository listed at (10), containing a curated list of various tools, to be used for data gathering. However, this list is finite and various other tools, that are preferred within your organisation, naturally can be used as well.
Depending on the needs and wants of your organisation there are specific sources that focus on certain types of news, for example: Where the RSS feeds of the National Cyber Security Centre (NCSC) excels in providing well-structured and documented data relevant to imminent or potential (operational) cyber threats, such as the announcement of a vulnerability in software, it will most likely lack in notifying you when there are geopolitical conflicts unfolding between the countries you might be dependent on, or operating in.
Combining the right sources and tooling (11) is of crucial importance to organisational data needs, with proper mapping of the Situational Awareness levels.

Step 3 Data processing & post analysis

Raw data needs to be processed before analysis. This involves tasks such as putting data into spreadsheets, decrypting files, translating foreign language data, and assessing data reliability. Pre-processing improves the value of data analysis by cleaning, normalising, and transforming the data, making it consistent and usable. This saves time and allows for structured, automated, and manual analysis of the most relevant data.

Finally, the data review is a critical component of Context Driven Data Gathering. After processing the data, a thorough analysis is needed to answer the questions mapped during the requirements phase. The team works to translate the dataset into useful recommendations for stakeholders. This iterative process requires constant maintenance and adjustments to stay aligned with the organisation's mission and vision across enterprise levels. Staying up-to-date with relevant data gathering is crucial.

Conclusion

The Context Driven Data Gathering Framework aims to reduce the amount of irrelevant data that analysts have to filter through, thereby reducing analyst fatigue and improving decision-making. By establishing and defining the context of the organisation and prioritizing data sources based on organisational needs, the framework can help organisations to collect higher quality and more relevant data, ultimately leading to more efficient and effective data analysis.

References

(1) Pherson, R.H. and Heuer, R.J. (2021) Structured Analytic Techniques for intelligence analysis. Thousand Oaks, CA: CQ Press.
(2) "Closing the Data Decision Gap" (2022), March. Available at: https://web-assets.domo.com/blog/wp-content/uploads/2022/03/Domo-The-DDG-Paper.pdf.
(3) Alahmadi, B.A., Axon, L. and Martinovic, I. (no date) 99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms. Oxford University. Available at: https://www.usenix.org/system/files/sec22summer_alahmadi.pdf
(4) Kabir, S. M. S. (2016). Methods Of Data Collection Basic Guidelines for Research: An Introductory Approach for All Disciplines (first ed., pp. 201-275).
(5) Schönberger, M. and Cukier, K. (2013) Big Data: A Revolution That Will Transform How We Live, Work, and Think. Houghton Mifflin Harcourt.
(6) Endsley, M.R. (no date) Designing for Situation Awareness in Complex System. SA Technologies, Inc. Available at: https://www.researchgate.net/profile/Mica-Endsley/publication/238653506_Designing_for_situation_awareness_in_complex_system/links/542b1ada0cf29bbc126a7f35/Designing-for-situation-awareness-in-complex-system.pdf
(7) Jakalan, A. H. M. A. D. (2013). Network security situational awareness. The Int. J. of Comp. Sci. and Commun. Secur. (IJCSCS), 3, 61-67
(8) Vijayan, J. (2022) DHS Review Board deems Log4j an 'endemic' cyber threat, Dark Reading. Available at: https://www.darkreading.com/application-security/dhs-review-board-deems-log4j-an-endemic-cyber-threat (Accessed: March 24, 2023).
(9) Inc., S. (no date) LOG4J updates and vulnerability resources, Sonatype. Available at: https://www.sonatype.com/resources/log4j-vulnerability-resource-center (Accessed: March 24, 2023).
(10) https://github.com/hslatman/awesome-threat-intelligence#tools
(11) https://github.com/hslatman/awesome-threat-intelligence#frameworks-and-platforms