Development of a sanitization tool to share securely private data for critical-infrastructure protection

FINAL DEGREE PROJECT
Student: Guillermo Aumatell Salom

Advisors: Jordi Forné Muñoz
and David Rebollo-Monedero

2018-2019
I would like to express my very great appreciation to Professors Jordi Forné and David Rebollo-Monedero, my research supervisors, for their guidance and encouragement.

I would also like to extend my thanks to the rest of the team for their help in offering me the resources.

Finally, I wish to thank my parents for their support throughout my studies.
Abstract

The anonymization of data is a relatively new and important aspect that allows to guarantee privacy and to be able to progress technologically without putting in danger confidential information. The project arose thanks to the participation of the Universitat Politècnica de Catalunya in the European project H2020, called CIPSEC. Its collaboration consists on the creation of a tool able to anonymize data picked up from cyber-attacks targeting different enterprises for its further analysis. This work consists on the creation of a tool which allows the sanitation of files for its further use being able to ensure the security of the information. Finally, this tool is a little platform which allows to perform a personalized sanitation.
Resum

La anonimització de dades és un aspecte relativament nou i important que permet assegurar la privatització i poder seguir avançant tecnològicament sense posar en risc informació confidencial. El projecte va sorgir gracies a la participació per part de la Universitat Politècnica de Catalunya en el projecte europeu H2020 denominat CIPSEC. La seva col·laboració consisteix en la creació de una eina que anonimitzi les dades recollides dels ciberatacs registrats en distintes companyies per el seu posterior ús en investigacions. Aquest treball consisteix en la creació d’una eina que permeti la sanitització de fitxers per el seu posterior ús podent assegurar la seguretat en la informació. Finalment aquesta eina és una petita plataforma que permet dur a terme una sanitització personalitzada.
Resumen

La anonimización de datos es un aspecto relativamente nuevo e importante que permite asegurar la privatización y poder seguir avanzando tecnológicamente sin poner en riesgo información confidencial. El proyecto surgió gracias a la participación por parte de la Universitat Politècnica de Catalunya en el proyecto europeo H2020 denominado CIPSEC. Su colaboración consiste en la creación de una herramienta que anonimice los datos recolectados de los ciberataques registrados en distintas compañías para su posterior uso en investigaciones. Este trabajo consiste en la creación de una herramienta que permite la sanitización de ficheros para su posterior uso pudiendo asegurar la seguridad en la información. Finalmente, esta herramienta es una pequeña plataforma con varios programas que permiten llevar a cabo una sanitización personalizada.
Revision history and approval record

<table>
<thead>
<tr>
<th>Revision</th>
<th>Date</th>
<th>Purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>01/12/2018</td>
<td>Document creation</td>
</tr>
<tr>
<td>1</td>
<td>27/12/2018</td>
<td>Document revision</td>
</tr>
<tr>
<td>2</td>
<td>04/01/2019</td>
<td>Document v0.9 finished</td>
</tr>
<tr>
<td>3</td>
<td>07/01/2019</td>
<td>Document Revision by supervisor</td>
</tr>
<tr>
<td>4</td>
<td>13/01/2019</td>
<td>Document v1. finished</td>
</tr>
</tbody>
</table>

DOCUMENT DISTRIBUTION LIST

<table>
<thead>
<tr>
<th>Name</th>
<th>e-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guillermo Aumatell Salom</td>
<td><a href="mailto:guillermoaumatell@gmail.com">guillermoaumatell@gmail.com</a></td>
</tr>
<tr>
<td>Jordi Forné Muñoz</td>
<td><a href="mailto:jforne@entel.upc.edu">jforne@entel.upc.edu</a></td>
</tr>
<tr>
<td>David Rebollo-Monedero</td>
<td><a href="mailto:david.reollo.m7@gmail.com">david.reollo.m7@gmail.com</a></td>
</tr>
</tbody>
</table>

Written by: Guillermo Aumatell  
Date 05/01/2019  
Name Guillermo Aumatell  
Position Project Author  
Reviewed and approved by: Jordi Forné  
Date 07/01/2019  
Name Jordi Forné  
Position Project Supervisor
Table of contents

Abstract ........................................................................................................................................... 3
Resum............................................................................................................................................... 4
Resumen.......................................................................................................................................... 5
Revision history and approval record ................................................................................................. 6
Table of contents................................................................................................................................. 7
List of figures ....................................................................................................................................... 8
1. Introduction ................................................................................................................................. 9
   1.1. Motivation .............................................................................................................................. 9
   1.2. Goals ...................................................................................................................................... 9
   1.3. Content .................................................................................................................................. 10
2. State of Art .................................................................................................................................... 11
   2.1. Sanitization for Cybersecurity .............................................................................................. 11
   2.2. CIPSEC ................................................................................................................................. 13
      2.2.1 UPC Contribution ........................................................................................................... 15
3. Tool Description ............................................................................................................................ 17
   3.1. Specifications ......................................................................................................................... 17
   3.2. Architecture .......................................................................................................................... 18
4. Sanitization (ANTeater) ................................................................................................................ 20
5. Policies (ANTpolicies) ................................................................................................................... 23
6. Audit (ANTanalyst) ....................................................................................................................... 27
7. Data Base ...................................................................................................................................... 28
8. Platform ........................................................................................................................................ 30
9. Conclusions .................................................................................................................................. 32
10. References .................................................................................................................................. 33
11. Annex .......................................................................................................................................... 34
    11.1. Code and Files .................................................................................................................... 34
    11.2. Interesting Regular Expressions ......................................................................................... 35
    11.3. CIPSEC File – Example ...................................................................................................... 36
List of figures

Figure 1. Data Security Lifecycle (DSL) .................................................................................. 12
Figure 2. CIPSEC - Project Overview .................................................................................... 13
Figure 3. CIPSEC - Partners .................................................................................................. 14
Figure 4. CIPSEC - UPC Contribution Architecture ............................................................... 15
Figure 5. CIPSEC - UPC Contribution Software .................................................................... 15
Figure 6. TheANT - DSL ....................................................................................................... 17
Figure 7. TheANT – DSL - 2 .................................................................................................. 17
Figure 8. TheANT - Architecture .......................................................................................... 18
Figure 9. ANTeater - Software................................................................................................. 20
Figure 10. ANTeater - example .............................................................................................. 21
Figure 11. ANTeater – example executed .............................................................................. 22
Figure 12. ANTeater – Result ................................................................................................. 22
Figure 13. ANTpolicies – Policy File ....................................................................................... 23
Figure 14. ANTpolicies – Policy File example ....................................................................... 24
Figure 15. ANTpolicies - Software ........................................................................................ 25
Figure 16. ANTpolicies – example ........................................................................................ 25
Figure 17. ANTpolicies – Predetermined options .................................................................. 26
Figure 18. ANTanalyst - Software ........................................................................................ 27
Figure 19. Data Base - Architecture ..................................................................................... 28
Figure 20. Data Base – ANT_LOGS example ....................................................................... 28
Figure 21. Data Base – ANT_ANON example ..................................................................... 29
Figure 22. Data Base – ANT_POL example ......................................................................... 29
Figure 23. Data Base – ANT_USERS example ..................................................................... 29
Figure 24. Platform - Connection ........................................................................................ 30
Figure 25. Platform – config.txt .......................................................................................... 30
Figure 26. Platform - Login .................................................................................................. 31
Figure 27. Platform - TheANT .............................................................................................. 31
Figure 28. TheANT – Repository ........................................................................................ 34
Figure 29. TheANT – Code .................................................................................................. 34
Figure 30. TheANT – Policies .............................................................................................. 34
1. Introduction

This final degree project deals with the development of a tool that allows the user to sanitize files in order to eliminate confidential or unwanted information. The whole tool has been created on the basis of the CIPSEC project, in which the Universitat Politècnica de Catalunya participates. Based on the demand for the project, the tool has been developed in a specific way trying to make the final product as complete as possible, not only to be used in cybersecurity sector, but in other areas too.

Throughout this document, the different parts of this tool will be studied and analyzed. The main reasons that led to take certain development decisions will be explained. And last but not least, several possible projections towards the future of this tool will be submitted.

Next, we present the reasons that have led to develop this tool.

1.1. Motivation

More and more data are generated in the world. All data analysts are eager to obtain them and be able to treat them in order to have even more objective information. This information can help to obtain great benefits. It can bring about better product propositions to the customers, higher sales, greater profits, and also better research or even higher estimates of future events. All sorts of benefits could be obtained if all these data were available, but it is not always possible. Much of this information is generated by users doing normal things in their lives like using the phone, studying, working, taking public transports... There are a lot of things that generate data, however we can’t use all this information! Why? The users do not want to relinquish it. Why should I share all my data with companies to allow them to make profits on my account? Why freely? Why should I share my data with all the people letting my preferences be known? If a company has a lot of personal data from their users, why can’t they use it to share and investigate? This is a matter of privacy! Nobody wants to give up to all the world his personal data. But what would happen if the collected information did not indicate any personal data? There would be no problem and information could be shared.

This project allows to anonymize the files that the user decides on the basis of certain conditions that will be seen later on. The files that will be anonymized contain cybersecurity data, attacks made to organizations. When those files are anonymized, they can be shared safely without transferring confidential information of the people who otherwise would be victims. This action permits to share information to improve the investigations.

1.2. Goals

The objective of this project is to have a tool that anonymizes files in a specific format to share data between organizations without any risk. That tool will be a small platform with three different software components that are going to be interconnected to reach better results. All those software can be independent from each other and each one has a specific function. The main goal is to reach a simple, clear and easy tool that anyone can use without sanitization knowledge nor expertise. It is meant to be like “a tool for Dummies!”
To reach this goal we use some technologies like Amazon Web Services, Anaconda (to use Python programming) and obviously my programming abilities.

1.3. Content

The content of this document is separated into three major parts: an explanation about the state of art in the world of data sanitization, a description of this platform and conclusions with possible projections towards the future. The state of art is an analysis of the current situation about data sanitization. I am going to present the CIPSEC project, the father of this project. The description of this platform contains five chapters, each one of the software that compose the platform. And finally the conclusion include some brief explanation of the improvements and a general overview of this project.

Welcome to The ANT Platform!
2. State of Art

Nowadays, data sanitization is a crucial challenge for any company; all the new technologic regulations should include information about this practice. The new EU General Data Protection Regulation (GDPR) [11] includes some clarifications about that practice and must be treated correctly to stick to the law. There are some examples that are actually present [7]:

- **Article 1**, section 17: Right to erasure
  Be able to prove any erasure of any data for audits.
- **Article 13**: Information to the Data Subject
  When the purpose for which data is collected expires, ensure that the data is available for erasure.
- **Article 39**: All inaccurate data should be securely erased and a record digitally signed must kept.
- **Chapter 3, Section 3, Article 19: Notification obligation regarding rectification or erasure of personal data or restriction of processing**
  Provide a certificate of erasure to the data subject.
- **Chapter 1, Section 5, Article 25: Data protection by design and by default**
  Be able to demonstrate that an end-of-life data is planned for.
- **Article 30: Records of Processing Activities**
  The certified erasure reports must include the data controller’s contact.

All that examples show us that data sanitization is present and must be considered. We will go into detail in the field that concerns us.

2.1. Sanitization for Cybersecurity

There is no regularization to sanitize files for the field of Cybersecurity, but of course some regulations exist that must be applied. All the files in the world can be different; not all the people use the same standards to save their files. Some examples could be Word, Json, PDF, PNG, txt, etc. We can think of this as the files that we have in our personal computer: we have a lot of files ordered (or not) in some folders and all of them are different. But all of them have a lifecycle that allows to treat and use the data that are stored. Most of the files are created and some time later are erased or forgotten in a dark and cold folder. That data progression is named Data Security Lifecycle (DSL) [8]. The DSL includes six phases from creation to destruction and it’s not a linear progression. Once the data is created, it can bounce between phases without any restriction. These are the different stages:
When a file is created, it is located in some storage system. But that’s not the actual utility of the file. We want to use that file, to share that file and to do something with it. Obviously to use and to share that information we need the owner’s authorization. But if we transform (sanitize) the data so that we cannot know who has generated it, we could share it and use it without problems because we aren’t giving any personal data.

Actually there are some different methods to sanitize data [1]. The techniques that we will comment are methods that can be applied in a lot of sources depending on the type of the source. Some of them will be better if we use them in a Database, others in a simple File and some of them even in Hardware.

- **Method: NULL’ing**

The first method is the NULL’ing. It can consist simply in deleting the entire data, replacing it with NULL value. In a database it could be a null value; that’s a term used to represent a missing value, a field that appears to be blank. In a file it could be “”, just a space bar. That technique is useful in certain occasions but not as the entire strategy to sanitize.

- **Method: Masking Data**

A second method is Masking Data. Masking data means replacing fields with a Mask Character such as ‘X’ or ‘A’ or even a word like ‘Person’… For example, a phone number could be like that:

93 894 55 55

And after applying the Masking Data method would appear as:

93 XXX XX XX

The masking characters remove some sensitive content from the phone number. If we are studying the percentage of calls that are made from a city, this sanitization method does not prevent us from analyzing it and we are hiding confidential information.

- **Method: Encryption/Decryption**

Obviously we can think about the encryption method. This technique offers the possibility to leave the data in a public space; only with the appropriate key the data will be visible. The big problem that we face is that the data cannot be used for possible analysis. And all the security depends on the strength of the encryption used.
• **Method: Substitution**

This technique is similar to the Masking Data with some interesting differences. This method consists of randomly replacing the content with data from a limited spool. That means that if we want to hide the address, we can substitute this information by another address from the same City. To do that we need a spool with some addresses from each city to be able to substitute them randomly. To generate the spool could be a major effort.

• **Method: k-Anonymous Microaggregation** [9]

This technique is a very interesting an innovative method to sanitize a set of data. It permits to sanitize a set of data distributing all this data in various groups to apply different methods in each one. This method offers excellent data utility in machine learning applications.

There are more data sanitization methods. We can think about a lot of different techniques, even a hardware physical destruction! These methods that have been named are just an example of some of the existing ones. To sanitize a Cybersecurity file, no guideline exists; it depends always on the needs of the users. Each field that needs to be sanitized requires an analysis to decide which transformation could be the best option.

2.2. CIPSEC

CIPSEC is the project that inspired the creation of this tool and allowed us to contribute to it. CIPSEC means: Critical Infrastructure Protection with innovative SECurity framework [2]. They are creating a unified security framework for Critical Infrastructures to orchestrate security products to offer protection in Information Technology and Operational Technology. The participation in this project on our part consisted in making the sanitization tool that will be presented in this framework. We are going to present the project to make it clear; and once finished, we will comment our participation and where our part is placed.

![Project Overview](image)

The main aim is to create a unified security framework. CIPSEC will offer a complete security system including services like vulnerability tests and recommendations, forensic analysis, standardization... There are thirteen partners that are collaborating to reach the objectives:
CIPSEC focuses on three Critical Infrastructures: health, environmental monitoring and transportation. These are the pilots of this project: three different CI environments.

- **Health**: Hospital Clínic de Barcelona (HCB)

The *Hospital Clínic de Barcelona* is one of the most recognized public tertiary University hospitals in Spain and in the EU. Some of the Operational Technology in the Hospital are: Access Control, Voice over IP, monitoring and therapeutic equipment... By analyzing some cyber-attacks to healthcare environments, it has been seen that they show an exponential increase. Stolen health credential are sold for $10 each, about 10 or 20 times the value of a credit card number. Studying some cases, we can find common patterns in cyber-attacks. However, we need more solid solutions to healthcare technological entities. That's why CIPSEC is working to improve these requirements.

- **Environment**: Network Air Quality Detection Regional System (AQDRS)

The *CSI Piemonte*’s network is composed by 78 fixed public stations, 6 mobile public stations, 900 sensors and it’s connected to several other IT systems (like apps and/or databases). CSI Piemonte is a company that gives IT services to the Public Administration. They have developed services from different sectors: health care, trade and industry, administrative systems... The objective is to improve the protection of the network.

- **Railway transportation**: DB NETZE

DB NETZE is the biggest business premise in Germany: 5700 stations as gates to railway transportation, more than 33km rail network... The objectives are to improve an innovative and future-oriented company, which is environmentally friendly and has a sustainable mobility, including a safety domain in all the infrastructure, thanks to this framework.
2.2.1 UPC Contribution

To reach this objectives CIPSEC is working on a Dashboard where all the partners will install their software. The Universitat Politècnica de Catalunya is working on the sanitization environment doing a tool to sanitize all the Data generated by the different devices located at the partners’ networks. It is a must for all these companies not to give any private information.

The desired architecture is the following:

![Figure 4. CIPSEC - UPC Contribution Architecture](image)

When an attack is received and the device has detected it, it sends a log to a Database. That database stores all this attack log under a single structure. That structure includes Source IP Address, Date, Users, Destination IP Hostname, Events, Devices, Organization and some others fields to extract the maximum amount of information possible...

But obviously all that Database is not sanitized and this is where we include the software.

![Figure 5. CIPSEC - UPC Contribution Software](image)

The results are saved in a database named MISP that means Malware Information Sharing Platform [12]. MISP is an open source software for collecting, storing, distributing and sharing cyber security data. The goal of MISP is to encourage the exchange of structured information within the security community. It provides functionalities to support the exchange of information. This software can work with different text files like CSV, JSON, XML, etc. The
software designed is prepared to receive the data with JSON format and to return the sanitized data with the same format.

If we think about these outlines, we can see that Data that are included in the MISP Database are sanitized. Therefore all that data don’t include any personal information. That means that we can share this final database without compromising anything.
3. Tool Description
We are going to present the Platform: The ANT. First of all, we will present the specifications and the needs that have led us to make the decisions that will be seen later. It must be taken into account that the basis of this project has been the European CIPSEC project and all decisions have been taken regarding its requirements.

3.1. Specifications
As indicated in the second section (2.1 Sanitization for Cybersecurity), there exists six phases in data progression. Where we participated was focusing on the Share phase. TheANT permits to share data without sensitive information.

If we change the direction of possible flow in the six phases, forcing the system to anonymize the information just before sharing it, we could promise total privacy by using these data. This means that there can be no danger in sharing this information.

Using these schemes, we built a platform that includes three software: ANTeater, ANTpolicies and ANTanalyst. All of them are used to sanitize file with JSON extension. Obviously they can be used independently of each other but they complement each other in order to get the maximum performance out of the tool.

In the following sections, each of the tools that make up this platform will be explained in detail. Now we are going to introduce them in order to contextualize their utilities.
Who carries out the anonymization is ANTeater. It is responsible for generating the sanitization and deliver to the user, if desired, the “clean” file. The user can choose the type of anonymization that he wants in each field of the selected file. Some types of anonymization are indicated in section 2.1 Sanitization for Cybersecurity. How can the user know exactly what is going to be applied? In order to know what is going to be applied, some policy files are used. This policy files are JSon files that contain all the necessary information. The anonymization of this project is based on the use of regular expressions: a method used in programming for pattern matching. So the policy files include the regular expressions needed and some more information that will be explained in section 5. Policies (ANTpolicies). To apply sanitization policies without difficulties you can use ANTpolicies. Finally, as stated in the second section (2. State of Art), GDPR includes some clarifications about how to prove any erasure of the data. To be able to adapt to the law, we are going to use a Database to save all the user actions logs. All this information can be analyzed and observed from ANTanalyst.

That has been a brief summary of the different tools and how they connect with each other. All this platform has been programmed with Python Language [5] due to CIPSEC Project requirements. Now let’s discuss how the architecture has been assembled to fit the pieces of the puzzle.

3.2. Architecture

TheANT Platform is a platform with three software that could be used to sanitize files and save logs. To do that it was important to have an architecture that treats the programs independently. That permits to simplify future improvements and even third parties can develop their programs and include them in a quick and simple way to the platform.

There are three blocks in this architecture: the software that can be used to sanitize files, the directory where the files (policy files and files to sanitize) are saved and the Database which is used to register all the logs that generate all the software.

Figure 8. TheANT - Architecture
Each software use the database to save every action that the user executes: sanitize a file, change a policy file, add a new policy to a policy file, which policy are used to sanitize a file... All this things are logged in the database to be able to keep track of the files. Other feature of this database is the possibility to have users that share this database. We will talk about that in section 7. Data Base.

This platform can be used with or without a Data Base Connection. It was created thinking in the possibility that a company could use it with several persons at the same time. If we only need to sanitize private files, it is possible to use it in local mode without any connection.

We are going to look in more depth at each of the aspects of this platform.
4. Sanitization (ANTeater)

ANTeater is the main software component of the platform. It’s the main mechanism that allows the user to sanitize a file and record it, if desired. We are going to see all the features.

First, we can see that it has a simple graphic interface without too many options. It’s designed exclusively to sanitize files. We can see a menu, two blank boxes and a button ‘Start’. On the left box we will see the file that will be sanitized and on the right box the sanitized file. Let’s go to see an example while we comment each feature.

If we display the menu, we can see the two main options: Load Document and Load Policy. We are going to see an example with a fake file to simplify reading. If we load a document \((\text{LOG\_00001.json})\) and a Policy File \((\text{policy\_logs.json})\) we will see that the document is displayed on the left box, and the options that are available on the policy file are displayed between the two boxes.
We can see in the file some sensitive data like IP Address (SRC_IP), the Operating System (OS, that’s a sensitive field due to Zero Day attacks), the Router ID (it could be sensitive due to the extra information that a user can obtain) and some others fields. We assume that we are interested to share this log with someone but we don’t want him to know our IP Address and our Operating System. We have selected a Policy File and it appears two options to sanitize: OS and Source IP. Each one can have multiple options to use. In this case we have selected a Partial Suppression sanitization for the field ‘OS’ and a Pseudonymisation option for the Source IP field. Each option will be explained in section 5. Policies (ANTpolicies).

If we click on Start, the result appears with the fields that we were interested on sanitizing. The field ‘SRC_IP’ after the execution indicates 255.255.255.255. We can know that it is an Address IP, but not the value. And the OS is indicating us that the Operating System is Windows, but we don’t know the version. Then we are protected from some attacks (or at least hindering their job).
Finally, if we agree with the result, we can extract the file and save it in the Repository. As we had mentioned, this platform works well with a company with several people using this platform. So, to speed up the saving process, the final file is saved with a name using a pattern including the date of the execution.

We can see that it includes first the word RESULT and finally the date in which it has been executed.

In conclusion, to sanitize a file we need to have prepared a Policy File. To create it we can use any text editor (Notepad, Notepad++, jedit, Left...) or we can use ANTpolicies that facilitates widely our work.
5. Policies (ANTpolicies)

ANTpolicies is the software that helps to create or modify policy files. The policy files have a simple structure meant to be simple and clear. You don’t need to use this tool to modify each policy file but it can help you. We are going to study the two main elements of this tool: the files and his functionalities.

All the policy files have the same structure:

```json
{
  "policy": {
    "NAME": "POLICY NAME 1",
    "NUMBER": "0", // Not necessary but it can helps in some functionalities
    "TYPE": {
      "ACTION": "SUB POLICY NAME 1",
      "CONTEXT": "NOT USED, BUT IT CAN BE INTERESTING",
      "REGEX": "REGULAR EXPRESSION",
      "REPLACEMENT": "SUBSTITUTION IN REGULAR EXPRESSION",
      "TARGET": "TARGET OF THE REGEX"
    }
  },
  "policy": {
    "NAME": "POLICY NAME 2",
    "NUMBER": "0", // Not necessary but it can helps in some functionalities
    "TYPE": {
      "ACTION": "SUB POLICY NAME 1",
      "CONTEXT": "NOT USED, BUT IT CAN BE INTERESTING",
      "REGEX": "REGULAR EXPRESSION",
      "REPLACEMENT": "SUBSTITUTION IN REGULAR EXPRESSION",
      "TARGET": "TARGET OF THE REGEX"
    }
  }
}
```

Figure 13. ANTpolicies – Policy File

Each file has some policies, each one with the same fields. In each policy there may be several sub-policies indicating the name, the regular expression, the target of the regular expression and the substitution to use in the regular expression.

What is a Regular Expression? This project is not aimed to explain in depth the regular expressions by themselves, but it’s true that they are the basis of the functioning of the tools. A regular expression [6] (regex or regexp) is a sequence of character that defines a searching pattern. We are going to show a simple example to understand it better.
Let’s imagine that we have the following regular expression: ‘([A-Z])\w+'. That regexp means it will match all the set of characters that contain an uppercase letter [A-Z] and some characters after it \w+. We are going to apply this regular expression to any text:

**Regular Expression:** ((A-Z)\w+)

**Text:** “In a village of La Mancha, the name of which I have no desire to call to mind, there lived not long since one of those gentlemen that keep a lance in the lance-rack, an old buckler, a lean hack, and a greyhound for coursing. An olla of rather more beef than mutton, a salad on most nights, scraps on Saturdays, lentils on Fridays, and a pigeon or so extra on Sundays, made away with three-quarters of his income.”

We can see that the regular expression matches all the words starting with an uppercase which have some characters after it. This is the operation of regular expressions; with a defined grammar you can look for patterns in texts and modify them.

As we have seen, the policy files are based on regular expressions and now that we know better the operation of a regular expression we are going to analyze the policy file that we used in the example of section 4. Sanitization (ANTeater).

Figure 14. ANTpolicies – Policy File example

---

1 Cervantes Saavedra, M. (1608), The Ingenious Nobleman Sir Quixote of La Mancha.
As we can see, we have two main policies: Source IP and OS. The first one, Operating System, is searching a content that includes some letters followed by a number. This is useful for windows versioning. This regular expression serves to mask the windows version. The second policy: Source IP is searching an IP Address. Depending on which sub-policy you choose, it transforms data into one or other solution. For example, a Partial Suppression transforms an IP Address 194.178.1.1 into 194.178.X.X. Pseudonymisation policy transforms data disabling all the contents and showing 255.255.255.255. These are just some example of a policy file. Let’s see how the tool works.

That’s the main aspect of this tool. As we can see, this tool serves to modify policy files without needing to know the structure we have seen at the beginning of this section.

We are going to Load the policy file that we are studying: ‘policy_logs.json’. When the policy file is loaded, we can see all the content of the file on the tool and, of course, it can be modified.

We can change the content of the file and press Save to obviously save the changes that we have made. The New button resets all the fields and the Delete Button deletes the Policy that is
checked. If we are a user that doesn’t know too much about regular expressions, some helps exist. There are some predetermined options that can be chosen and include the information.

![Figure 17. ANTpolicies – Predetermined options](image)

In conclusion, the ANTpolicies Tool facilitate the creation of policy files even without knowing anything about regular expressions. Once we have a policy file created, we can execute the sanitization with ANTeater (Section 4. Sanitization (ANTeater)).

Finally, the last tool of this platform allows to audit all this actions. Obviously ANTeater and ANTpolicies generate some data, of each action that is executed, for example: by who, where, when, etc. That means that it’s possible to control all the actions that the users executed with this Platform. This last software is ANTanalyst.
6. Audit (ANTanalyst)

The last software is ANTanalyst, a software that serves to Audit all the actions that have been executed with the ANT platform. It is a useful element since it allows the visualization of the Database in an orderly manner and without having knowledge in SQL.

The tool has two labels that allow to enter the necessary information: a date and the application. The date could be a single day in the format DD/MM/YY or even the word ‘All’ to see all the historical logs. The field application permits: ‘Eater’, ‘Policies’ or ‘Info’.

Eater option shows the files and the policies files used to sanitize files. Policies option shows the modifications did it on each policy file. Finally, Info option shows a general information about the platform during the date selected.

In this example we can see that we choose the application Eater on a selected date. The information that concerns ANTeater appears showing us all the actions executed in this date.

The operation of this application lies on the use of SQL language. Requests are sent to the platform server and the results are displayed in an easily readable way for someone who does not master these technologies. The internal functioning of the database will be explained in the following section (7. Data Base) where the structure of the code is included.

Obviously the use of this tool lies in the ability of the user in front of Oracle databases. In case of knowing this technology previously, it is preferable to use visualization tools such as SQL Developer before using this tool. What has been sought is to be able to offer an extra so as not to need previous knowledge in any of the fields except for the interest in computer security.
7. Data Base

The ANT platform database is a relational database with star scheme. The star scheme consists in one fact table (the main table) referencing any number of dimension tables. The dimension tables contain the real data where we can find more information and the fact table contains the minimum information to be able to find data in dimension tables.

When we talk about relational database we are talking about Keys. In each table we have some columns and obviously some of them are most important than others. Some of them contain the Primary Key: the columns that define every register. We can’t have two registers with the same primary Key in a single table. When we talk about a Foreign Key we are saying that the columns that have this Foreign Key are related to columns with Primary Keys from other tables.

In our architecture we have the primary Key implanted in ANT_LOGS table and each of the others tables contain a Foreign Key with the same value.

![Figure 19. Data Base - Architecture](image)

We are going to explain each entity (table) from our scheme to understand it better.

- **ANT_LOGS**: It contains a general description from an event: the user that executed the action, the status of the action, the date in which he did it... All this information is named Metadata. That’s not the real data, only some of general data which is always generated regardless of the origin. We are going to take the register that has been created while we were executing the example of this document.

  ![Figure 20. Data Base – ANT_LOGS example](image)

- **ANT_ANON**: It contains the real data about ANTeater tool. We can appreciate that the ID is the same that in ANT_LOGS table. That’s because ANT_LOGS table is the fact table (the main table of the scheme) and this column allows to establish a relation between the two tables. The rest of the information contains real data that can be exploited by business to get relevant data about the use of the tool.
**ANT_POL**: It contains the real data about ANT policies. We can see that it also has the field ID to have a relation with the fact table ANT_LOGS. The rest of the information contains the changes that a policy file suffer. That permits to know each variation of the files and have a register of it.

As you can see, the database has a very simple structure. So searching for information in it is not complicated. The code that is executed for each software is divided in two different packages in the Database: “Analyst” and “Logs”. The first package, the Analyst’s one, is in charge of getting the desirable information to show in the ANTanalyst tool. The second package, Log’s one, is in charge of getting data from all the actions that generate the user using the tools and save it in the database.
8. Platform

Finally, we are going to see how all these elements are joined into a single software. We can create a shortcut to execute the software and the first frame that appears is the following one:

![Platform - Connection](image)

*Figure 24. Platform - Connection*

The platform offers the connection to the Database which was described previously. Actually these configurations are enabled in a file named: ‘config.txt’.

![config.txt](image)

*Figure 25. Platform – config.txt*

This file can be changed to configure others connections. The file to clone the database architecture is in the annexes. Once the connection has been chosen, the main frame appears to login or sign up in the platform. The two buttons on the lower right corner are simply advertising to visit the official social networks of TheANT Platform.
Once you have login into the platform, you can choose to execute each software.
9. Conclusions

Finally, the project that consisted in creating a tool that would allow to anonymize files with a specific format was transformed into a more functional and not so closed tool. We managed to make it accept any JSON file with different formats, and make it able to sanitize the chosen fields, using a clear and simple pattern which is based on a plain text file. It works not only with computer security documents but with any data. The needs of the user define the way in which a file will be anonymized using this powerful tool: regular expressions.

Difficulties were always present, when faced with the generalization so that it could work with any file. A lot of time was needed to make it as simple as possible, both functionally and technically. The user that will harness this software doesn’t need to know anything about programming! What we sought at all times was to satisfy current needs of the European CIPSEC project, yet not to remain there but go further and reach more areas with the same software. Thanks to all the difficulties, this platform was born.

Of course it can keep on improving and growing far more, combining different technologies. If I was to improve this work, the first improvement I would love to implement would be the use of machine learning for automatic detection of data type and then automatic sanitization, automatic email system to help to improve the platform, database encryption... For anonymization is one of the fundamental ethical challenges of nowadays in the information society.
10. References


11. Annex

11.1. Code and Files
This project is composed of a tool and default policy files for initial use. All this material is attached to this project and this is the repository:

![Repository](image)

*Figure 28. TheANT – Repository*

In each folder we can find the documents that we need to execute the platform.

The code is composed by the next files:

![Code](image)

*Figure 29. TheANT – Code*

In the Policies folder we can find some default policy files.

![Policies](image)

*Figure 30. TheANT – Policies*

All files sanitized are placed on the Results folder. On the document folder the user can place his files.
### 11.2. Interesting Regular Expressions

<table>
<thead>
<tr>
<th>REGEX</th>
<th>REPEX</th>
<th>TARGET</th>
<th>COMMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\d{1,3}).\d{1,3}.\d{1,3}.\d{1,3}$</td>
<td>1\d{1}2X.X</td>
<td>IP Address</td>
<td>Subnet Mask /16</td>
</tr>
<tr>
<td>(\d{1,3}).\d{1,3}.\d{1,3}.\d{1,3}$</td>
<td>255.255.255.255</td>
<td>IP Address</td>
<td>No information</td>
</tr>
<tr>
<td>192.168.\d{1,3}.\d{1,3}</td>
<td>192.168.X.X</td>
<td>IP Address</td>
<td>Only privates address</td>
</tr>
<tr>
<td>10.(5[3-9]</td>
<td>7-9).\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3}.\d{1,3} Well Known Port</td>
<td>Ports</td>
<td>Ports between 1-1023</td>
</tr>
<tr>
<td>^([0-9]</td>
<td>[1-9][0-9]</td>
<td>1[0-9][0-9]</td>
<td>1[0-9][0-9][0-9]</td>
</tr>
<tr>
<td>^([A-Z][a-z]+)/\d{1,2}*</td>
<td>1 Version Unknown</td>
<td>Operating System</td>
<td>Operating System without the version</td>
</tr>
<tr>
<td>^00::40::96</td>
<td>0:</td>
<td>:w</td>
<td>w</td>
</tr>
<tr>
<td>^7C:C3:A1:0:</td>
<td>:w</td>
<td>w</td>
<td>:1</td>
</tr>
</tbody>
</table>
11.3. CIPSEC File – Example

{"AlarmEvent": {
    "USERNAME": "",
    "SRC_IP": "45.77.86.114",
    "BACKLOG_ID": "563102ff7123414eb55fa6f60219da36",
    "DATE": "2018-07-17 18:42:07",
    "DST_IP": "10.244.67.117",
    "USERDATA7": "",
    "USERDATA6": "",
    "FILENAME": "",
    "PRIORITY": 4,
    "RELIABILITY": 10,
    "ORGANIZATION": "clinic",
    "SENSOR": "AD14C6F3975ED9860E32190EA3DF2535",
    "SID_NAME": "directive_event: Detected access to SAMBA in Honeypot",
    "USERDATA2": "",
    "USERDATA3": "",
    "USERDATA1": "tcp",
    "PROTOCOL": 6,
    "RISK": 4,
    "USERDATA4": "",
    "USERDATA5": "",
    "EVENT_ID": "2a99be5ab6ee43588c83b17029420b",
    "USERDATAB": "",
    "USERDATA9": "",
    "PLUGIN_NAME": "cyber-monitor",
    "DST_IP_HOSTNAME": "00000000",
    "RELATED_EVENTS": ":[89f011e885a4080027e052c6df02d06,89f11e885a4080027e052c1b3cd914],
    "PASSWORD": "",
    "PLUGIN_SID": "2",
    "CATEGORY": "Recon",
    "SRC_IP_HOSTNAME": "00000000",
    "SUBCATEGORY": "Scanner",
    "RELATED_EVENTS_INFO": {
        "a": {
            "userdata2": null,
        }
    }
}}
"date": "1531852927",
"plugin_sid": 14,
"src_port": 51332,
"log": "lkp1bCAxNyAyMD00MjowNyAxMCw4LjEuMIA/Pj8/Pz9kaW9uYWVhLXN5c2xyZyb2NhcHRlZCB3b3luZWN0aW9uc2VhcmVvX2hvc3Q9NDUuNzcuODYuMTE0LHJlbV9wb3J0PTUxMzMyLGNvbW5zcG9ydD10Y3AsY29uX3Byb3RvPXN1YmQl:",
"event_id": "89f111e885a4080027ea052c1b3cd914",
"filename": null,
"src_ip": "45.77.86.114",
"device": "10.0.2.44",
"plugin_id": 110000,
"username": null,
"tzone": "2.0",
"dst_ip": "84.88.67.117",
"interface": "enp0s3",
"dst_port": 445,
"fdate": "2018-07-17 18:42:07",
"organization": "clinic"
}
}