The Cyber Intelligence Analyst's Cookbook

Volume 1: A primer for Open Source Intelligence Collection and Applied Research

Open Source Researchers
THE OPEN SOURCE RESEARCH SOCIETY

This book is dedicated to my loving wife, children, and household spirits. My pursuit of knowledge has put them through a lot over the years and continues to do so. I'd also like to dedicate this book to the leaders I've worked closely with over the past few years. Without their support I would not have come this far. I'd also like to thank all the analysts that I've met over the years. Their passion and support for this industry is truly inspiring. Lastly, I'd like to attribute part of this dedication to all the professors, who over the years, have taken the time out of their day to deal with my inquisitiveness.

Table of Contents

Preface

This book. Well, it started out as a manual, or rather a brain dump of my process. I've spent the last year or so examining how I collect Open Source Intelligence (OSINT) and tag it. Pretty simple right? Not so much. What I found over that year was that I continually added new tags to the artifacts, or I was creating new tags because they didn't exist within the database I use for storing this information. I use the Malware Information Sharing Platform (MISP) exclusively for my work. MISP is open, expandable, and can be queried by other apps using several different methods. Most of all, it's free.

Anyways, I started with this brain dump of my process for recording OSINT. The work initially started out just for me. I haven't documented any of my methods, thoughts, what have you in quite some time. I was due for this knowledge transfer. However, as I began writing, I found that a manual wasn't going to cut it. The next thing I know, I'm writing a book, and thirty days-ish later, the first draft was completed. Truthfully, it's an awful book, and I apologize to anyone who attempts to read it. Yet, as I look back over the body of knowledge, I see that I've at least created a good foundation for future volumes. Opportunities for expansion and clarification. Who knows, maybe someone will find what's in this book useful.

The book itself is explicitly written for cyber intelligence analysts. Still, anyone who performs intelligence as a discipline can deconstruct what's here and apply it to any intelligence domain. I'm also assuming the reader, at a minimum, has access to the Internet and can look up the tools used within the book. I've tried my best to add references to the right level of detail and completeness. I do believe in citing sources. Well, I've been beaten into always citing sources through my academic career as a student. So, what exactly is in this book? Part 1 of this book goes over the way I collect and store OSINT into MISP. Part 2 goes over some higher-order analysis that can be applied to the data.

I've placed the book under the GNU Free Documentation License. I've learned a lot from the open community and feel that this particular contribution belongs to the community. Those who take part in the open community, per se, made me. I've had to put a lot of work into myself to get to this point of knowledge in my own life, but I would not have gotten to this point if others hadn't laid the foundation before me. I'm sure folks will argue with the premises and processes I've laid out in this book, and that's totally cool with me. Hell, the one thing I know from my current Ph.D. program at university is to be prepared for the beating. This book is in no way a stone tablet or bible that must be adhered to as gospel truth.

Open Source Researchers

Part 1 - Collection

Chapter 1: General Overview of OSINT Artifact Recording Process

Chapter 1 Introduction

This chapter covers the basic workflow for the recording of Open Source Intelligence (OSINT) using the Open Source Research Society's (OSRS) methodology for research[1]. The goal of the OSRS collection methodology is to capture information in a way that allows for higher-order analysis of the data. Most of the time, many organizations consider the work of intelligence complete when the info is ingested in an automated fashion. It is not to say that automated ingestion of information is not essential; it merely does not provide any meaningful analysis of the data. In higher orders of study, we seek to apply academic research rigor to the data, meaning the data itself is not only seen as informational but structured to become variables that contain layers of information if the right research methods are applied. Some examples of higher-order quantitative research methods are regression, correlation, ANOVA, and Chi-Square. These methods are important because they take the analyst to a level of statistical analysis of the data that extends beyond the use of descriptive statistics.

Nevertheless, the industry of Cyber Intelligence has not evolved (or matured) into this level of analysis, whereby the majority uses it on a day-to-day basis. With this in mind, the OSRS has created the Comprehensive Modular Cybersecurity Framework (CMCF) as a tool that is used in conjunction with the Malware Information Sharing Platform (MISP). Using the CMCF in conjunction with MISP, an analyst can create categories of variables regarding intelligence information that can then allow for the application of higher-order analysis of the data. The goal of the CMCF is simple: create a diverse enough taxonomy system that sufficiently describes the information with parallel context found within the various data types in cybersecurity. The CMCF is not an entirely new set of languages. Instead, the CMCF is a compilation of industry-standard frameworks like HITRUST, MITRE, the Cyber Killchain, and the Common Weakness Enumeration (CWE) framework. The CMCF also allows an analyst to not only use the CMCF alongside these industry-recognized frameworks but to augment their work when these frameworks by themselves do not sufficiently describe the data.

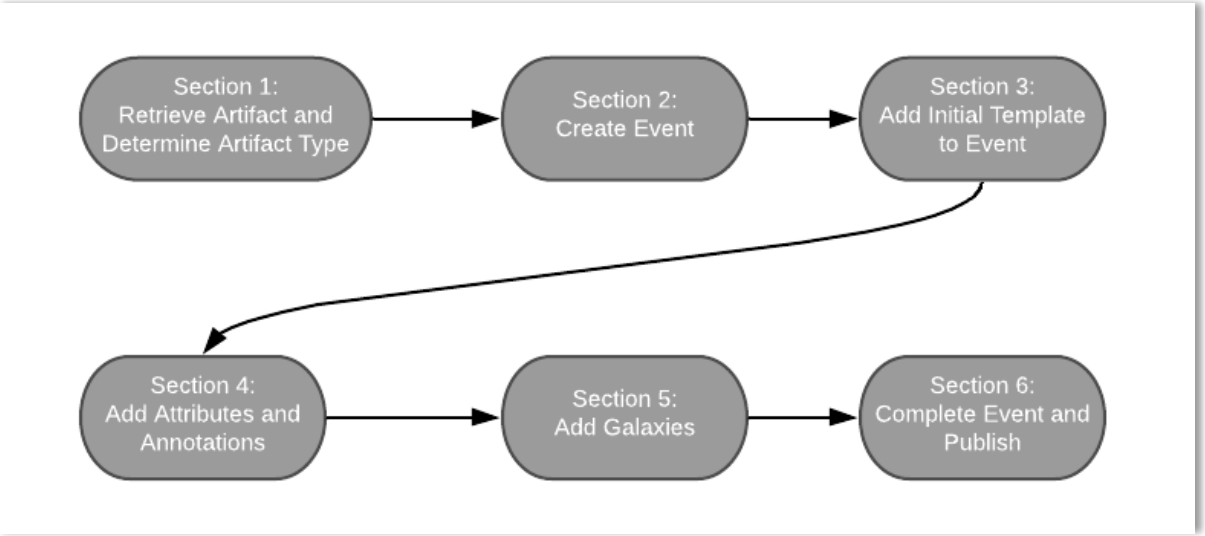
General Process for Recording OSINT Artifacts

The general process for recording OSINT artifacts follows an orderly and organized set of phases or sections. There are several reasons for this. The first reason is training. By standardizing the process, new analyst training becomes easy, as well as having their training measured against the standard. The second reason for standardization is for the normalization of the categorical placement of the data within a database.

Moreover, this type of standard categorization places the data into containers that form the basis for variable organization used during the application of higher-order analysis. Lastly, by standardizing the process, we can perform process reviews, make adjustments, and apply various management maturity models that place the process in a repetitive cycle of continuous

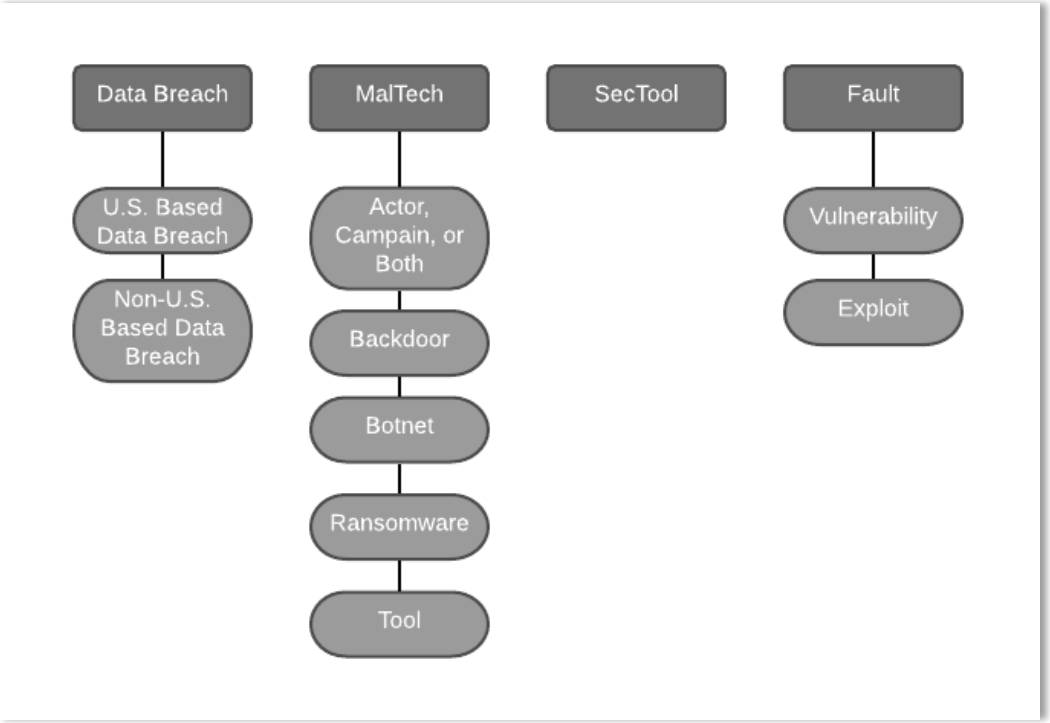
[1]. This resource is an OSRS specific resource and is supplemental to the CIRL primary MISP user and administration guide that found at the following URL: https://www.circl.lu/doc/misp/

improvement. The following diagram gives a visual representation to the general process flow for recording OSINT.



Section 1: Retrieve Artifact and Determine Artifact Type

This section covers the retrieval of a typical OSINT artifact. Within the scope of this book, an OSINT artifact is not a direct feed of specific indicators. Instead, artifacts are sets of data collected from publicly available web sites that offer a journalistic approach to presenting the information. Typically, an analyst will have a tool or method for retrieving these types of artifacts. Some examples of tools include RSS Feeds, Free Form Web Search, and Web Scrapers. It is a good practice to organize the information by subject to make the extraction of artifact context easier. Typically, these types of OSINT artifacts are organized into a hierarchical taxonomy. At the top of the hierarchy are the parent categories with their corresponding child categories below. The following diagram shows this hierarchy.

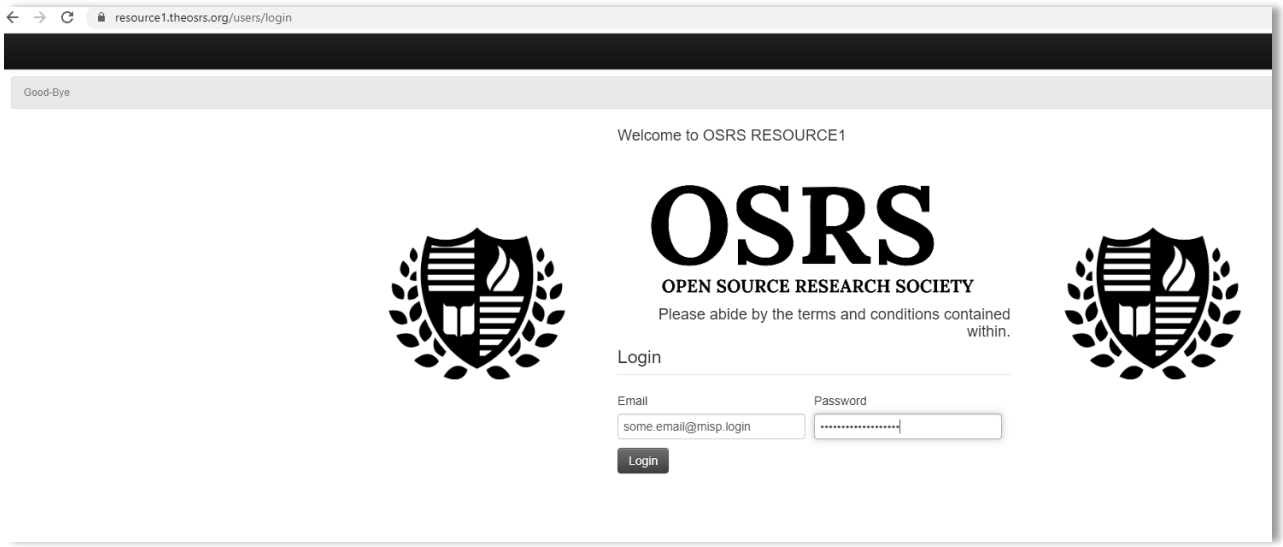


The parent categories determine the general concept of the artifact. The Data Breach category is for any artifact whereby the CIA triangle has been violated, regardless of the causality of the event. The MalTech category is for artifacts that have something to do with malicious software (i.e., Ransomware), threat actors, or campaigns. The SecTool category is for artifacts whose information relates to tools that are used by cybersecurity professionals for various reasons. Lastly, the Fault category is for classifying artifacts that contain information that relates to a known vulnerability or published exploit code.

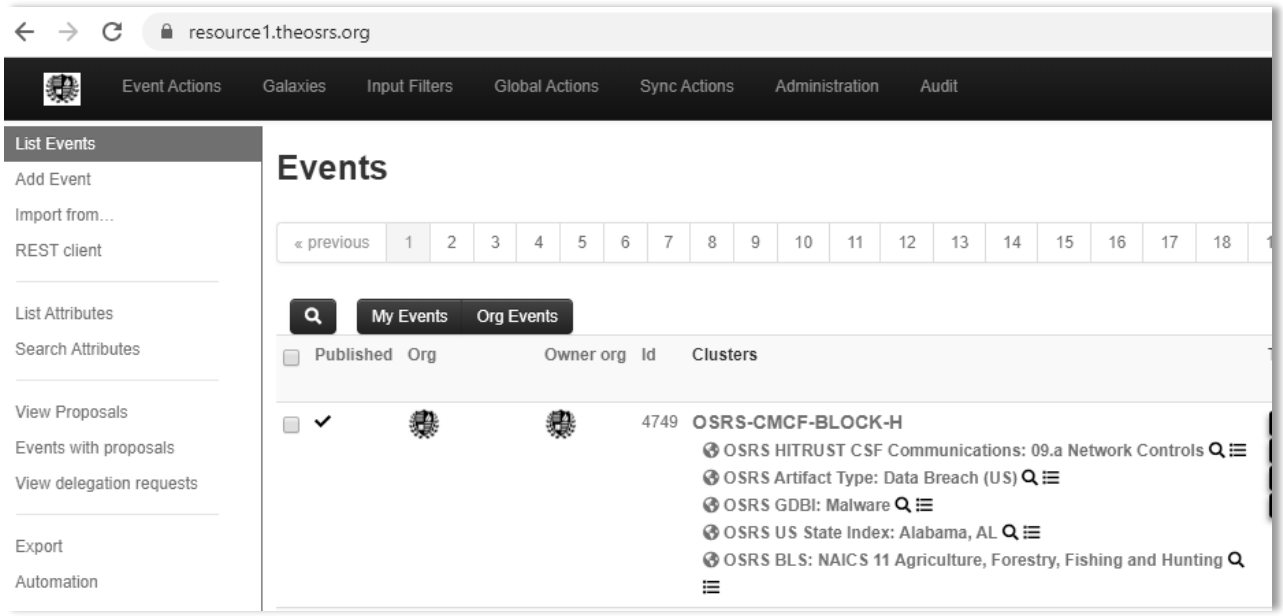
The frequency by which artifacts are collected and entered into MISP is determined by the individual analyst's availability and expertise. Both of these factors contribute to the speed and accuracy for which the information is entered into MISP. Either way, the artifacts will never leave the aforementioned hierarchical ontology of categories.

Section 2: Create Event

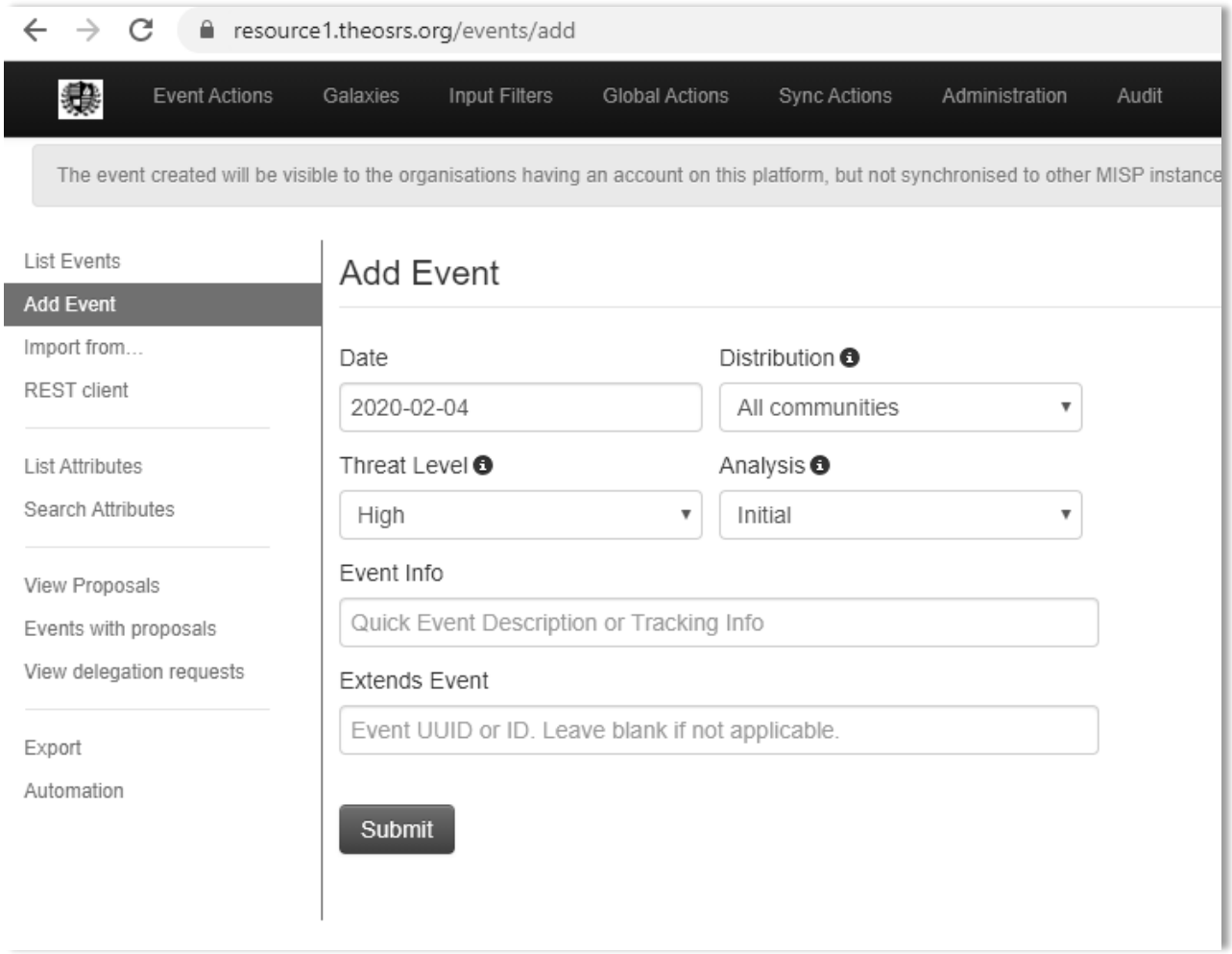
Once the artifact type is determined, it is time to create the event in MISP. The first step into creating the event is to log into MISP. We will navigate to the appropriate URL to access the login page for MISP. For this section, we will use the OSRS URL for MISP.



As seen in the image, the URL is entered into the search bar. After hitting *enter*, the page resolves to the MISP homepage, where we type in our credentials. MISP uses an email address format for usernames and a password scheme determined by the administrators. At this point, we will enter our credentials and log into MISP.



In the upper left corner, we will notice there is a link to *Add Event.* Clicking this link takes us to the initial event creation page.



After arriving at this page, we notice a set of data points that require completion in order to create the event successfully. Typically, an analyst will adhere to a standardized set of procedures during the event creation process set by their organization.

Date

The date of the event should reflect the date for which the event was created. No other date information from the artifact should be entered here. For example, the date for which the author of the artifact published the artifact publicly is not an appropriate use of date during event creation.

Distribution

All information contained within the OSRS MISP instance, RESOURCE1, is considered TLP: WHITE, and therefore should be available to all connected communities. Thus, the *All Communities* selection must be set for this particular section of event creation. The level of distribution may be modified by the organization and affects the level of delivery.

Threat Level

Though the image shows the threat level selected to be rated at *High*, doctrinally, the OSRS does not believe the analyst is positioned well enough to determine the actual level of threat as there are many factors at the individual organization level that can move this ordinal rating. Thus, all events, unless deemed explicitly by a quorum of peers, is set to *Undefined*.

Analysis

This section must be changed to a status of *Completed.* Once the initial information has been entered into MISP, there should be no need to go back to the event. However, there is one caveat to the completeness of an event. This caveat is the result of incompleteness in the built-in tagging system, which will be explained later on in this chapter. If, while completing data entry for the event, an analyst finds that there are no relevant tags relating to the information from the artifact, the event would then be tagged with an ontology that describes the incompleteness within the workflow. This ontology is provided by CIRCL/MISP and is an inherent capability of the tool, which also will be discussed in later chapters.

Event Info

This section is for giving the event an actual working title. Like other sections of the event creation page, the OSRS follows a standardized procedure for naming events. The following examples show how each type of event is titled by artifact type.

1) Data Breach
    a. Data Breach U.S.
        i. Syntax: OSINT Data Breach U.S.: <title from artifact>
        ii. Example: OSINT Data Breach U.S.: Massive PHI Breach at Local Hospital
    b. Data Breach Non-U.S.
        i. Syntax: OSINT Data Breach Non-U.S.: <title from artifact>
        ii. Example: OSINT Data Breach Non-U.S.: Massive PHI Breach at Local Hospital
2) MalTech
    a. Actor, Campaign, or Both
        i. Actor
            1. Syntax: OSINT Threat Actor: <title from artifact>
            2. Example: OSINT Threat Actor: Mr. Fox Strikes Again
        ii. Campaign
            1. Syntax: OSINT Campaign: <title from artifact >
            2. Example: OSINT Campaign: Orange Peel Botnet Campaign
        iii. Both
            1. Syntax: OSINT Actor/Campaign: <title from artifact>
            2. Example: OSINT Actor/Campaign: Mr. Fox and the Blue Box Malware Campaign
    b. Backdoor

        i. Syntax: OSINT Backdoor: <title from artifact>
        ii. Example: OSINT Backdoor: Purple Cheese Backdoor Strikes Windows
    c. Botnet
        i. Syntax: OSINT Botnet: <title from artifact>
        ii. Example: OSINT Botnet: The Cool Kid Botnet Takes Over Gaming Machines
    d. Ransomware
        i. Syntax: OSINT Ransomware: <title from artifact>
        ii. Example: OSINT Ransomware: Lock Your Stuff Ransomware Strikes Again
    e. Tool
        i. Syntax: OSINT Tool: <title from artifact>
        ii. Example: OSINT Tool: Book Remote Access Trojan on the Rise in Asia
3) SecTool
    a. Syntax: OSINT SecTool: <title from artifact>
    b. Example: OSINT SecTool: SQLMap Updates
4) Fault
    a. Vulnerability
        i. Syntax: OSINT Vulnerability: <title from artifact>
        ii. Example: OSINT Vulnerability: CVE-2020-12345 Windows Remote Code Vulnerability
    b. Exploit
        i. Syntax: OSINT Exploit-DB: <title from artifact>
        ii. Example: OSINT Exploit-DB: Jira Use-After-Free

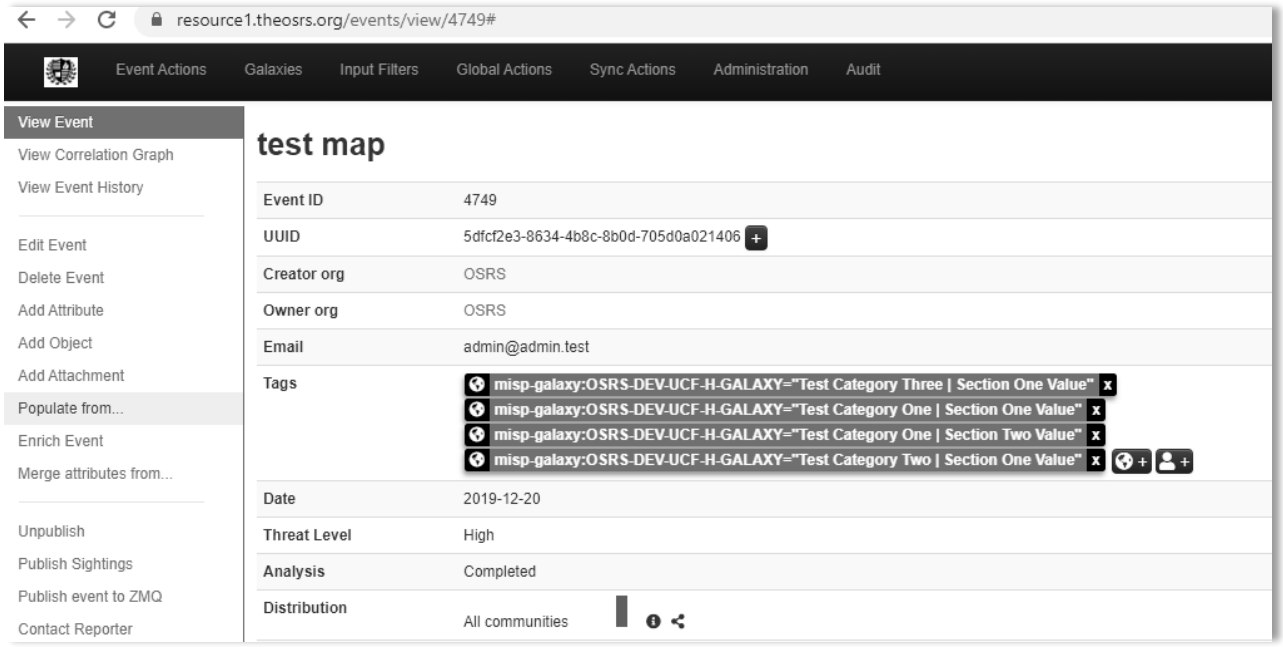
Extends Event

This section is not used by OSRS analysts during event creation and must remain blank.

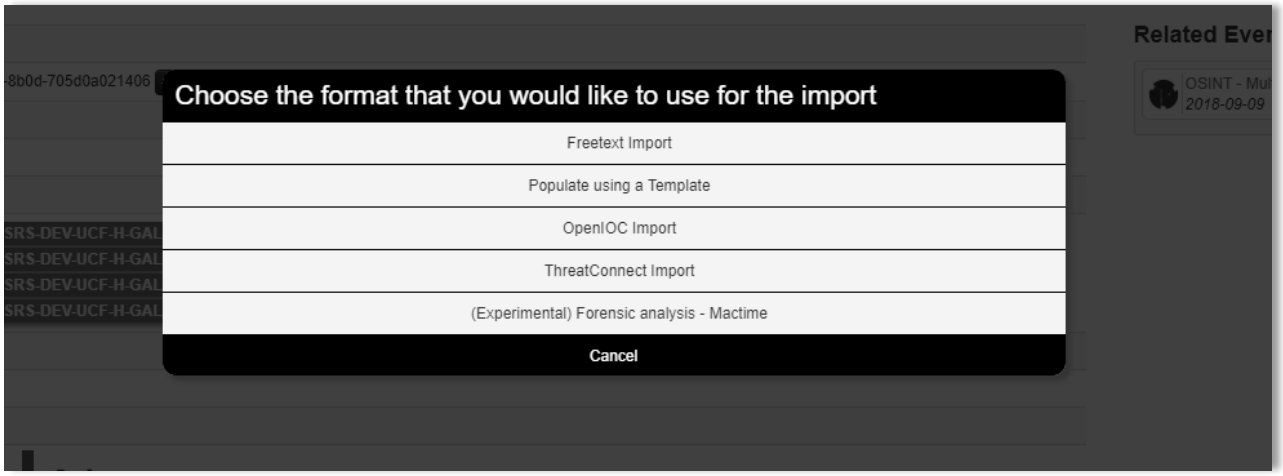
Once all the relevant information has is entered correctly, we click the submit button. By clicking submit, the event is added to the database.

Section 3: Add Initial Template to Event

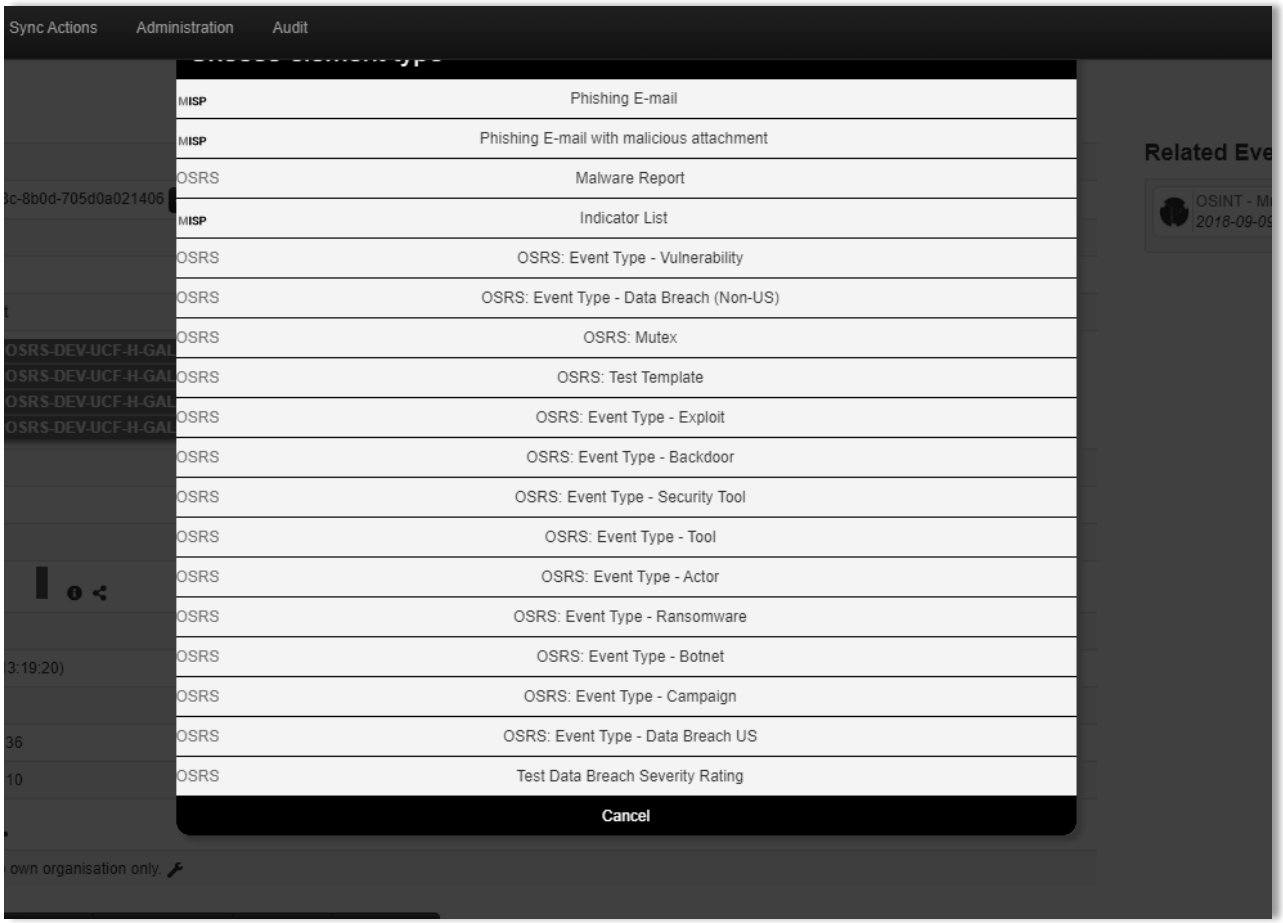
After the event is successfully created, we will apply the initial template. The purpose of this template is to capture important information about the event. Typically, the data collected by the template becomes the first set of attributes and tags associated with the artifact. The OSRS maintains several templates that align to the system of ontological artifact categories. To add information via a template, we first select the link *Populate from…* which is located to the left side of the event interface, as seen below.



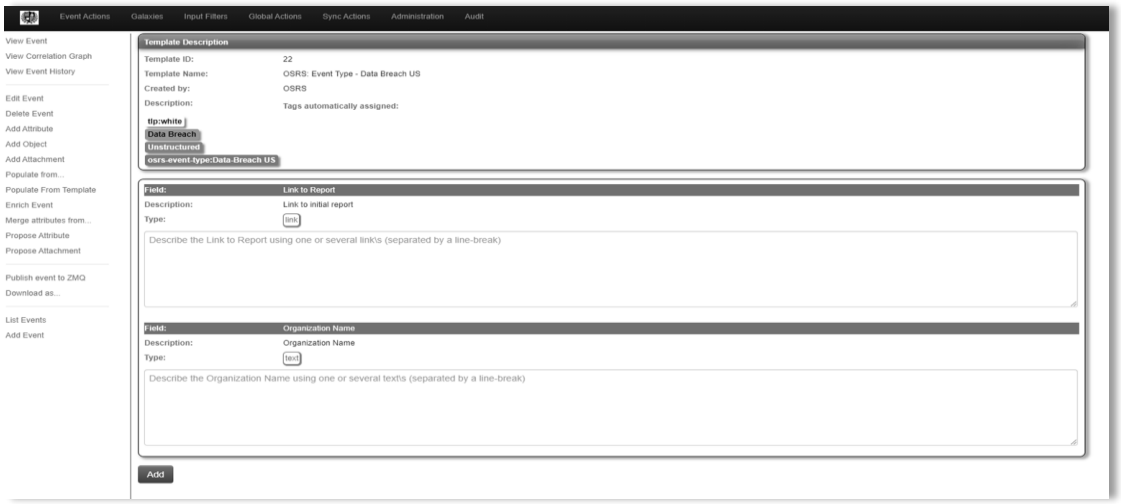
After selecting the *Populate from…* link, we are presented with a set of options.



At this point, we will select the option to *Populate using a Template.* This selection will bring up a further set of options. This set of possibilities lists the actual templates that will be used by an analyst. Below is an example of the templates available for an analyst to use when adding the initial information to an event.



For example's sake, we will select the template titled *OSRS: Event Type – Data Breach US.* By simply clicking on the text, we are taken to the templates input page where the analyst will enter in the relevant information.



As seen in the image, we can add the hyper-link of the original artifact to the report and the name of the organization(s) affected in the data breach event. Once we type in (or paste) the relevant information, we click the *Add* button to add the information to the event.

The following templates are available by artifact subject and the information that is captured by the template:

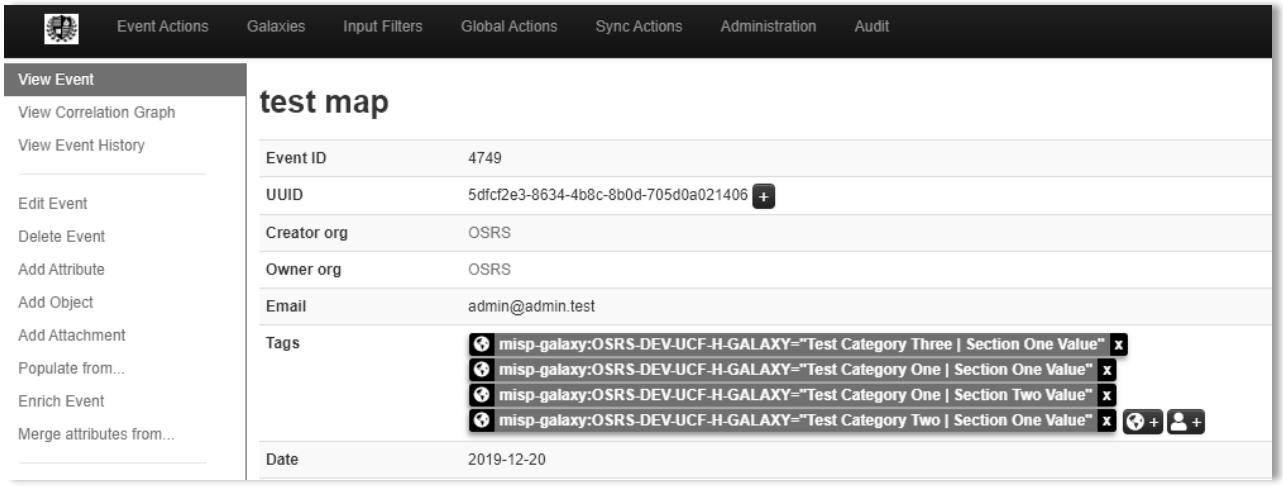
1) OSRS: Event Type – Data Breach US
    a. Link to the original artifact
    b. Organizations affected by the breach
2) OSRS: Event Type – Data Breach Non-US
    a. Link to the original artifact
    b. Organizations affected by the breach
3) OSRS: Event Type – Campaign
    a. Link to the original artifact
4) OSRS: Event Type – Actor
    a. Link to the original artifact
5) OSRS: Event Type – Backdoor
    a. Link to the original artifact
6) OSRS: Event Type – Botnet
    a. Link to the original artifact
7) OSRS: Event Type – Ransomware
    a. Link to the original artifact
8) OSRS: Event Type – Tool
    a. Link to the original artifact
9) OSRS: Event Type – Security Tool
    a. Link to the original artifact
10) OSRS: Event Type – Vulnerability
    a. Link to the original artifact
    b. CVE
11) OSRS: Event Type – Exploit
    a. Link to the original artifact
    b. CVE

CIRCL/MISP also maintains a set of templates that can be used during event creation or afterward in order to add attributes to an event quickly. These templates are out of scope for this book but can be found by using the same methodology for adding initial information to an event using OSRS templates. To see what information is captured in these additional templates, simply attempt to add them to an event to see what types of information they are useful for recording.

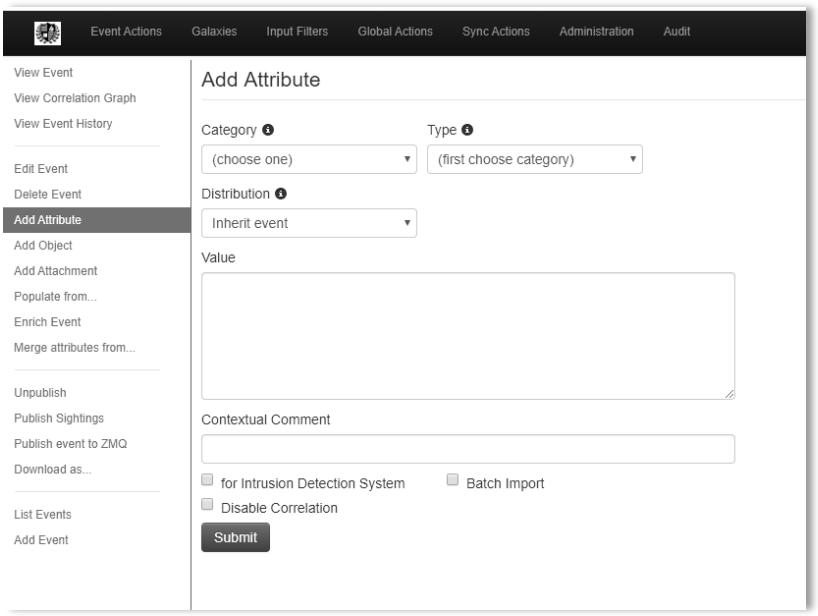
Section 4: Add Attributes and Annotation

After the initial information is added to the event, we are ready to record the rest of the data found within the artifact. Adding attributes can be achieved in several ways. For the purpose of

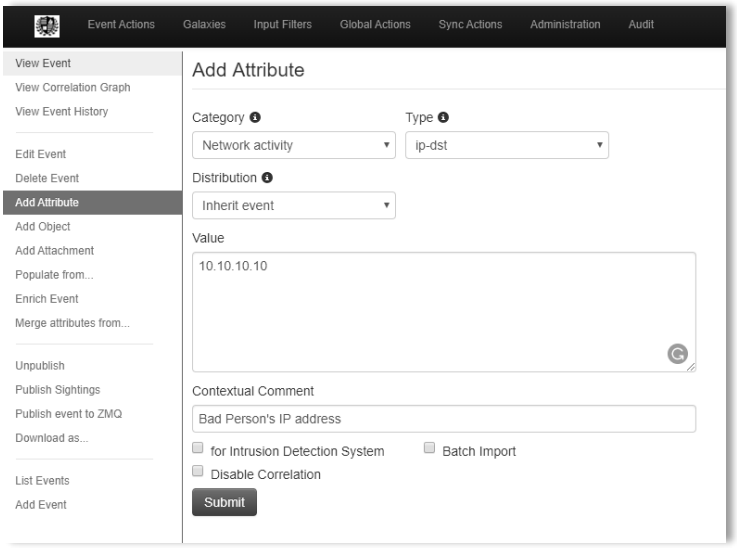
this book, we will consider adding attributes individually or by a template. To add individual attributes, we will click the link *Add Attribute* found on the left-hand side of the main event page.



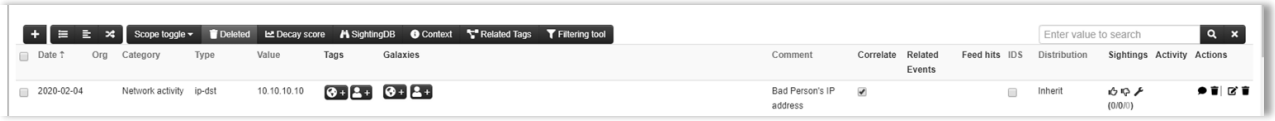
Once we click the *Add Attribute* link, we will be presented with a page that is very similar to the one used during event creation.



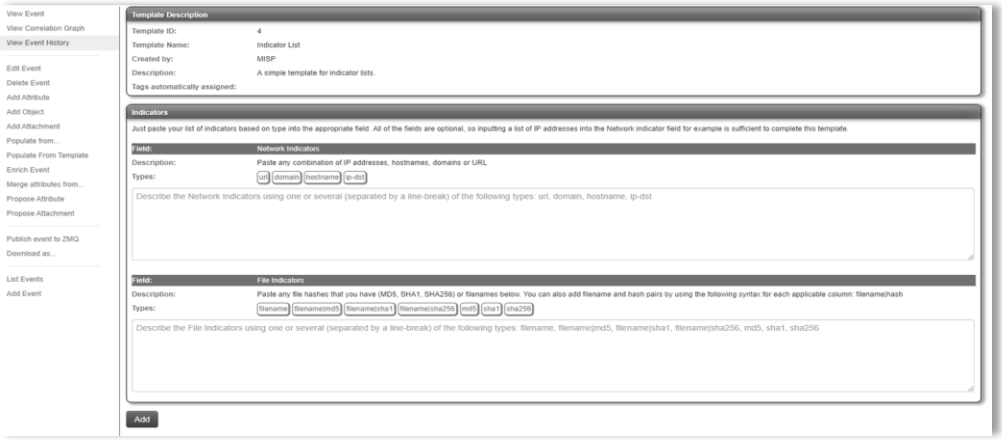
Several attribute types can be found within a single artifact. Some of the most common are IP addresses, file hash, and text relating to the topic of the artifact. To add the individual attribute to the event, we must first select the proper category and type. Below is an image of an example showing how we would add a destination IP address to an event.



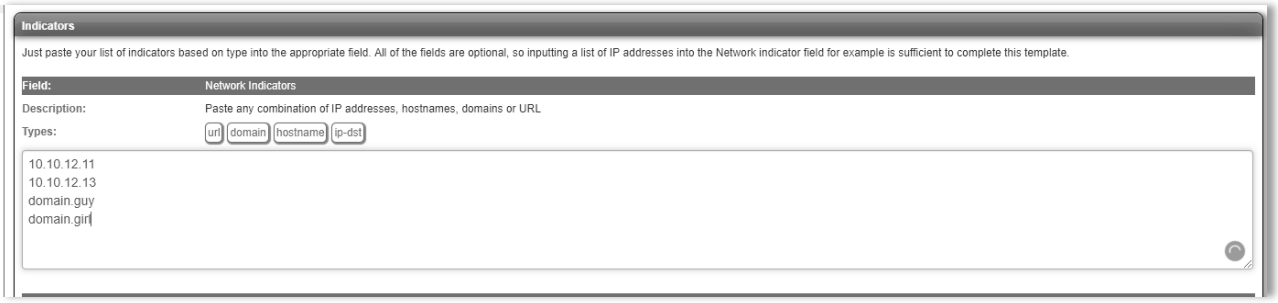
In this example, the *Category* has been set to *Network Activity*. The type has been set to *ip-dst* to represent a destination IP address. We then manually entered the IP address of *10.10.10.10,* which in this case, as manually annotated in the *Contextual Comment* section, is the *Bad Person's IP address.* Once all relevant flags and information are set, we will click the submit button to add the information to the event.



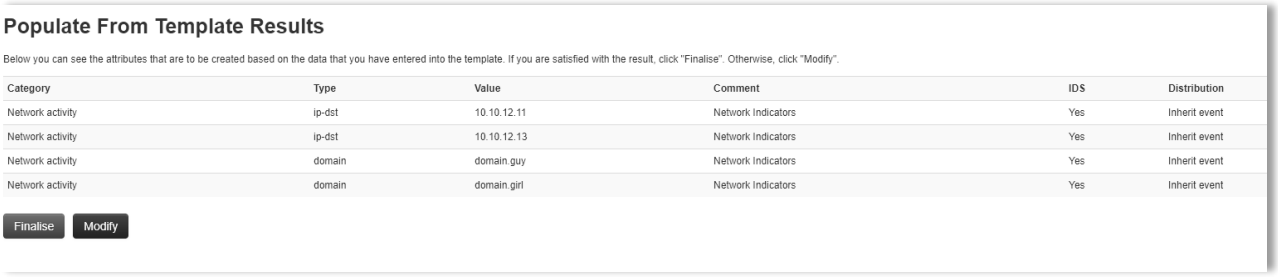
After clicking submit, we can verify that the information has been added to the event. The other way to add information to an event is by using the various templates. For this example, we will add several IP addresses using the *Indicator List* template provided by CIRCL/MISP. Using the steps previously outlined for adding information by template, we will navigate to the *Indicator List* template page. Once there, we will see a template that looks like this:



This template allows us to enter in bulk amounts of information, and we can see that the data entered can be varied by type. We will now enter a few IP addresses and domains.



Notice that the indicators in the above image are not the same as the previous IP address that previously entered. If we attempt to add in redundant information, even if relevant, MISP will return an error to us, letting us know that somewhere in the list is a duplicate attribute. We will not be able to add the attributes to the event until the redundancy has been corrected. Once the data entry has is completed, we will click the *Add* button located below the data entry section of the template to add the attributes to the event. However, there is one last step before the attributes are added. MISP templates give us the ability to review the information before adding the attributes to the event. Below is an image of what it looks like to an analyst during the attribute review portion of the template operation.



Once we are confident of the results, we can click the *Finalize* button, which will add the attributes to the event. If not satisfied in the results, we can click the *Modify* button, which will take us back one step and return us to the data entry portion of the template process.

Like the event creation section, the OSRS maintains a set of standards for recording information from an artifact. In most cases this will include, but is not limited to the following information as attributes:

1) Text Information
2) Network Information
    a. IP address
    b. URL
    c. Domain
    d. Host Name
3) File Information

a. Hash
b. File Name

There are other types of information that can be collected, such as the author of the artifact, the malware name, and the source of the artifact. In the case of data breaches, we will record the generalized state information where the data breach took place (i.e., Alaska, Michigan, and Texas). This type of information is captured during the tagging procedure of the artifact recording process. There are various doctrinal and procedural reasons for this information to be recorded as a galaxy tag rather than an attribute. However, those reasons fall outside the scope of this book. After all the relevant attribute information has been entered from the artifact, we can move on to the next section of this process, which is to annotate the artifact.

Some artifacts are simple to record. They offer easy to find information that the analyst can scribe into the event without little effort to remember what type of information is presented within the artifact. However, not all artifacts are the same. They vary in degree of context and complexity. Not all authors perceive the art of information dissemination the same. Therefore, there are gaps in the overall consistency in the presentation of artifact information. Ultimately, each artifact must be scrutinized by the analyst while reviewing and recording the data. As explained in *Section 5: Add Galaxies*, the use of annotation during the review and recording process becomes fundamental in the types of galaxy tags added to the event. The CMCF was created as a way to help analysts remember the possible ways in which an author may (or may not) portray information in an artifact. This framework works categorically and aligns to the hierarchical categories of ontology used to categorize artifacts. Some of the CMCF categories include industry-relevant frameworks like the MITRE CWE. Examples of annotation are outside the scope of this chapter. They will be explained in follow-up chapters that take us through a few real-world examples of how we might record specific artifact types.

Section 5: Add Galaxies

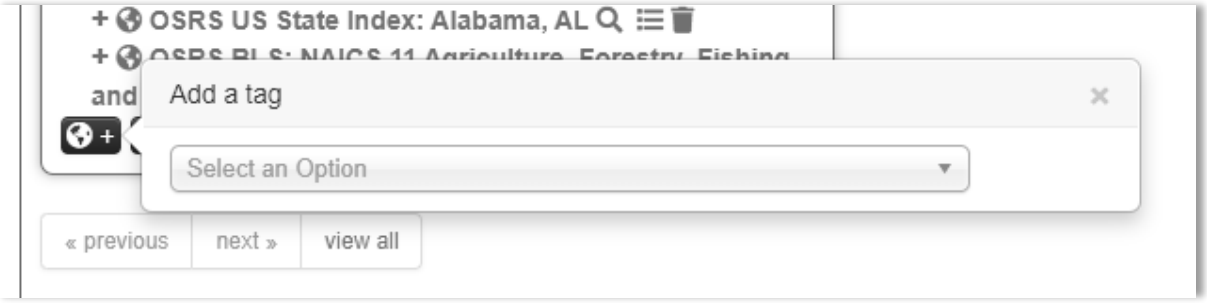
Galaxies are the elements of the MISP taxonomy system built within the application. MISP itself comes with a set of galaxies, but analysts can also create their own set of galaxy tags if need be. After the attributes have been added to the event, the relevant galaxy tags are added. Like event creation, the OSRS maintains a set of standards for tagging artifact events. The following table uses the CMCF and shows the types of tags associated with each type of artifact by artifact category. Across the top of the table are the general categories. The first column of the table is for the relevant CMCF sections.

	Data Breach	MalTech	SecTool	Fault
General Artifact	Artifact Type GDBI GECI US State Index ENISA SE DB Rating BLS Code	Artifact Type	Artifact Type	Artifact Type
Author/Source	OSINT Author OSINT Source	OSINT Author OSINT Source	OSINT Author OSINT Source	CVE Author Exploit Author Vulnerability Source OSINT Source
Block-H	Any item from the Block-H set of 16 categories	Any item from the Block-H set of 16 categories	Any item from the Block-H set of 16 categories	Any item from the Block-H set of 16 categories
Combined Ops Framework	Any item from the Combined Ops Framework's six categories	Any item from the Combined Ops Framework's six categories	Any item from the Combined Ops Framework's six categories	Any item from the Combined Ops Framework's six categories
MITRE Pre ATK	Any item from the MITRE Pre ATK 15 categories	Any item from the MITRE Pre ATK 15 categories	Any item from the MITRE Pre ATK 15 categories	Any item from the MITRE Pre ATK 15 categories
MITRE ATK	Any item from the MITRE ATK 12 categories	Any item from the MITRE ATK 12 categories	Any item from the MITRE ATK 12 categories	Any item from the MITRE ATK 12 categories
MalTech	Actors, Campaigns, or Both Backdoor Botnet Ransomware Tool	Actors, Campaigns, or Both Backdoor Botnet Ransomware Tool	Actors, Campaigns, or Both Backdoor Botnet Ransomware Tool	Actors, Campaigns, or Both Backdoor Botnet Ransomware Tool
CWE/Exploit DB	CWE Exploit DB Type Exploit DB Verification Exploit DB Platform	CWE Exploit DB Type Exploit DB Verification Exploit DB Platform	CWE Exploit DB Type Exploit DB Verification Exploit DB Platform	CWE Exploit DB Type Exploit DB Verification Exploit DB Platform
Informational Drawbacks	General Artifact Block-H Ops Framework MITRE Pre ATK MITRE ATK CWE	General Artifact Block-H Ops Framework MITRE Pre ATK MITRE ATK CWE	General Artifact Block-H Ops Framework MITRE Pre ATK MITRE ATK CWE	General Artifact Block-H Ops Framework MITRE Pre ATK MITRE ATK CWE

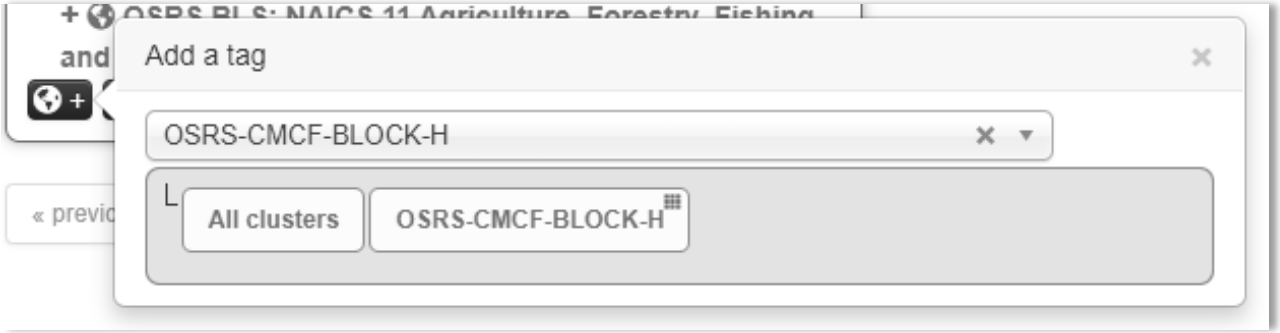
Obviously, by examining the table, not all of these tag categories will apply to every type of artifact. Additionally, there is a propensity due to the gaps in the way for which artifacts are reported that there will be more informational drawbacks than will be tags that describe the event. This topic of missing, vague, or the inclusion of information with no relevant tag becomes more apparent as we work through actual artifacts in the preceding chapters.

We have several options for adding a galaxy tag to an event. For this chapter, we will use the CMCF map to add galaxies to an event. The first step is to access the CMCF map. To do this, we simply need to click the button that looks like a globe located within the galaxy section of the event page.

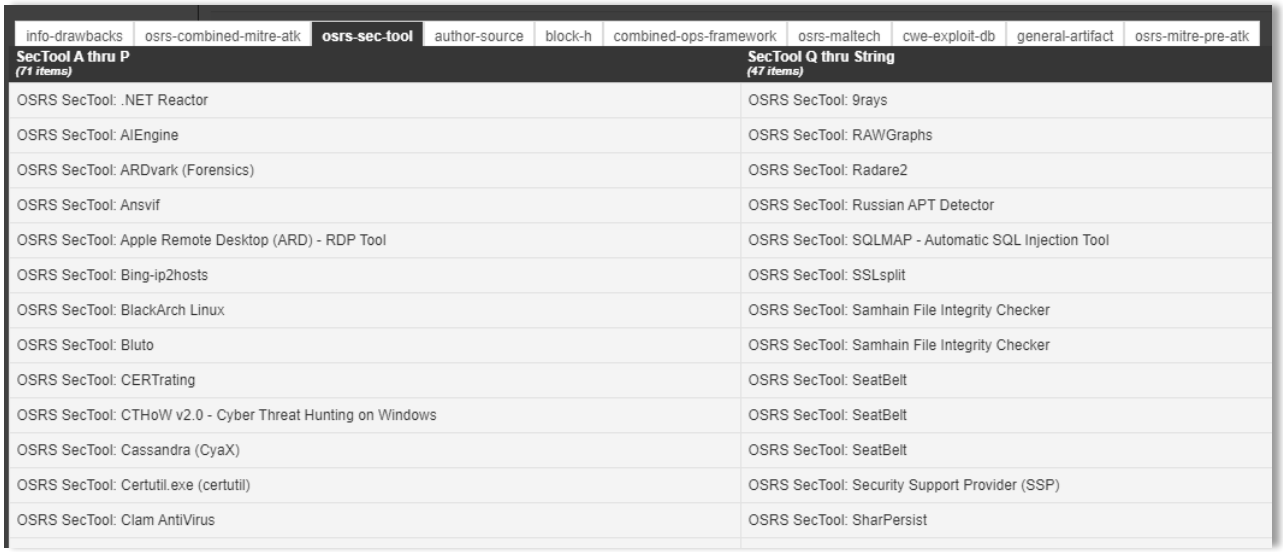
The globe we are looking for appears in the lower right corner of the galaxy portion of the event. Once clicked, we will be shown a list of items to choose from in the format of a drop-down menu.



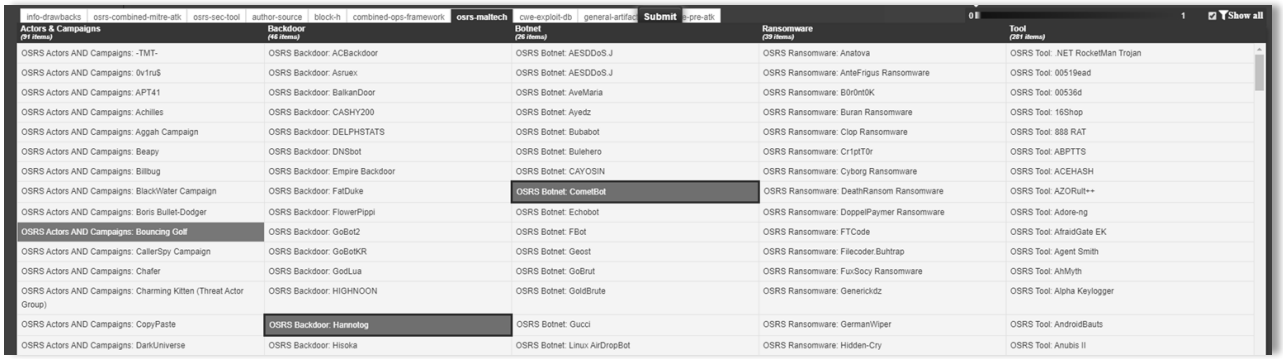
In this list, the analyst will navigate to and select the CMCF map. For this example, we will choose the box titled CMCF-BLOCK-OSRS.



After selecting the CMCF-BLOCK-OSRS option, we are presented with the CMCF map.



Once the map is presented to the analyst, there are two ways in which we can search for the appropriate tag. The first method is to find the relevant top-level category (i.e., author-source and osrs-maltech), click to that column, and then navigate through the options. Once the appropriate option has been found, we only need to click the option to have the option readied for submission. The convenience of this method is that an analyst can select multiple tags for submission at one time, thus reducing the extra process time it would take to place the individual tags. Another benefit is that the mapping also acts as an interactive encyclopedia. If an analyst is unsure of the selection, they only need to hover their mouse pointer over the option to get the textual definition of the tag. This encyclopedic quality helps to improve analyst accuracy by reducing the amount of doubt the analyst may have when applying the right tags to the event. The second method for searching and adding tags to an event is by using the search bar located at the bottom of the map, just above the *Submit* button. The search bar is interactive and begins searching for the relevant tag as soon as we start typing. Once a match has is found, we only need to hit enter on the appropriate tag from the list of presented options based on the textual search criteria. The following image shows tags that have been selected and readied using both methods of search selection.

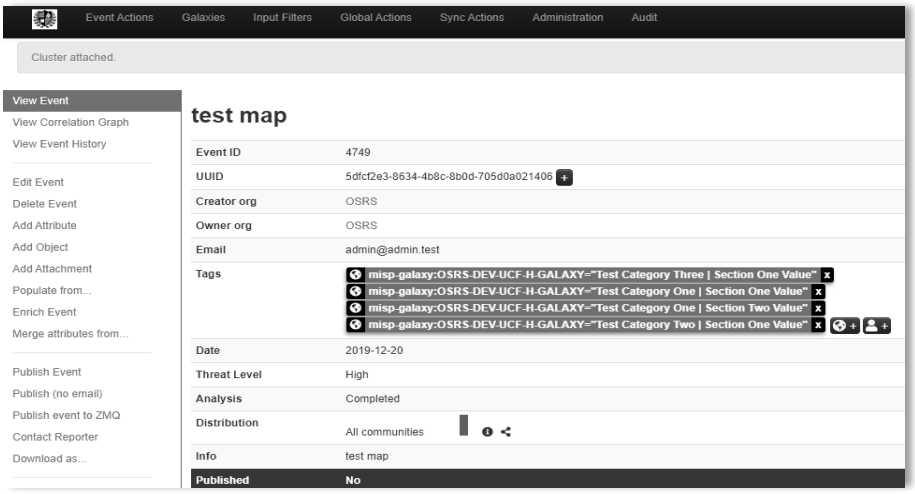


Once all tags are readied, we simply click the *Submit* button to have the tags added to the event. To verify the tags have been successfully added to the event, we look at the galaxy section of the

event to view the results of the submittal. Successful submission will show the tags now attached to the event.

Section 6: Complete Event and Publish

Finally, once we are confident that the information has been sufficiently recorded and annotated, the analyst can then publish the event. By publishing the event, we have notified users of the system that the event is complete. The process for publishing an event has two pathways. The first is to publish with email notification, and the second is without email notification. MISP has a built-in email delivery function that can be enabled by the system's administrator. If an event is published with email notification, an email will be created and sent to all users who have a legitimate email address used as their user account name. If the analyst selects the link to publish without email, no email notification is sent. The following image shows the location of the publish links to the left side of the event page.



Chapter Summary

In this chapter, we learned the basics of setting up a MISP event based on a collected OSINT artifact. In subsequent chapters, we will see how to synthesize this information and record the data for specific types of artifacts. It's important to understand that this methodology for recording OSINT is not only very labor intensive but not the only method. Ultimately, what this methodology does is record the data in a format that allows for further critical thinking, as well as data mining. This is very different than reactionary thinking. In cybersecurity, more specifically cybersecurity operations, there is a need for both critical and reactionary thinking. A cybersecurity operational environment is very fluid and dynamic. However, there are potential and kinetic relationships in decision making that can be recorded and studied. This methodology acts as an example to hopefully give others some form of inspiration when considering how they would blend both critical and reactionary thinking into their cybersecurity environments.

Chapter Acronyms

BLS: Bureau of Labor Statistics

CIA Triangle: Confidentiality Integrity Availability

CIRCL: Computer Incident Response Center Luxembourg

CMCF: Comprehensive Modular Cybersecurity Framework

CVE: Common Vulnerabilities and Exposures

CWE: Common Weakness Enumeration

ENISA: The European Union Agency for Cybersecurity

GDBI: Generic Data Breach Index

GECI: Government Entity Classification Index

HITRUST: Health Information Trust Alliance

MISP: The Malware Information Sharing Platform

OSINT: Open Source Intelligence

OSRS: The Open Source Research Society

URL: Uniform Resource Locator

Chapter References

CIRCL » CIRCL -- Computer Incident Response Center Luxembourg—CSIRT -- CERT. (n.d.). Retrieved February 4, 2020, from https://www.circl.lu/

Common Vulnerabilities and Exposures. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Vulnerabilities_and_Exposures&oldid=937348532

Common Weakness Enumeration. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Weakness_Enumeration&oldid=918897150

ENISA. (n.d.). Retrieved February 4, 2020, from https://www.enisa.europa.eu/

HITRUST. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=HITRUST&oldid=938925094

Home | U.S. Bureau of Labor Statistics. (n.d.). Retrieved February 4, 2020, from https://www.bls.gov/

Information security. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Information_security&oldid=939133610

MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing (formely known as Malware Information Sharing Platform). (n.d.). Retrieved February 4, 2020, from https://www.misp-project.org/

Open-source intelligence. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open-source_intelligence&oldid=937269341

URL. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=URL&oldid=930888655

Chapter 2: OSINT Artifact Recording - Data Breach Artifacts

Chapter 2 Introduction

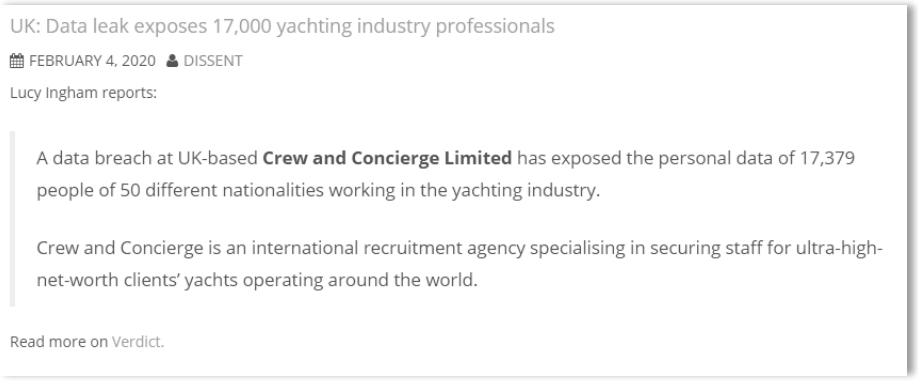
This chapter takes an analyst through the OSRS method for recording an OSINT data breach artifact. Like Chapter 1, we will be going through each step of the process and showing how to add the items to MISP. We will also be examining the artifact from the analyst's perspective and gaining insight as to *how* and *why* certain data is entered in particular ways. The goal of this chapter is to give analysts a baseline of exposure to the synthesis and reasoning behind the method for recording these types of OSINT artifacts.

Like Chapter 1, this chapter is organized into sections. Section 1 walks us through creating a data breach event. Subsequent sections 2 thru 10 show us how we would analyze the artifact to attach tags to the event. After the tags have been added to the event, we will score the data breach with a severity rating. We will also examine informational drawbacks and event annotation in greater detail.

Recording the Data Breach Artifact

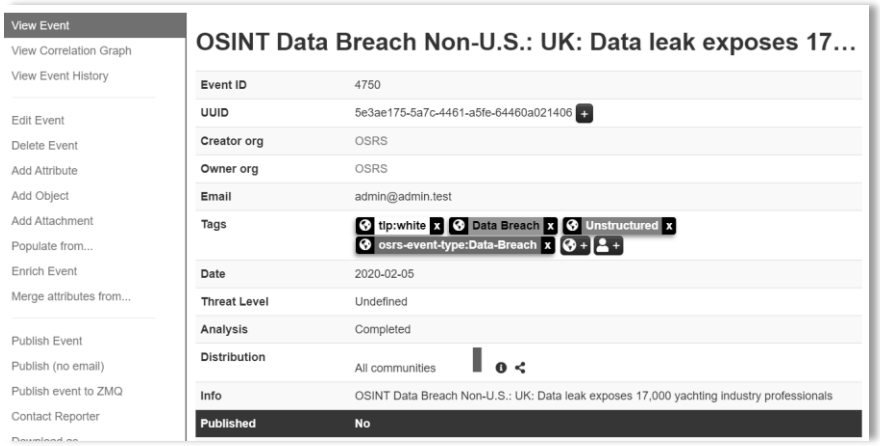
Section 1: Artifact Selection and Event Creation

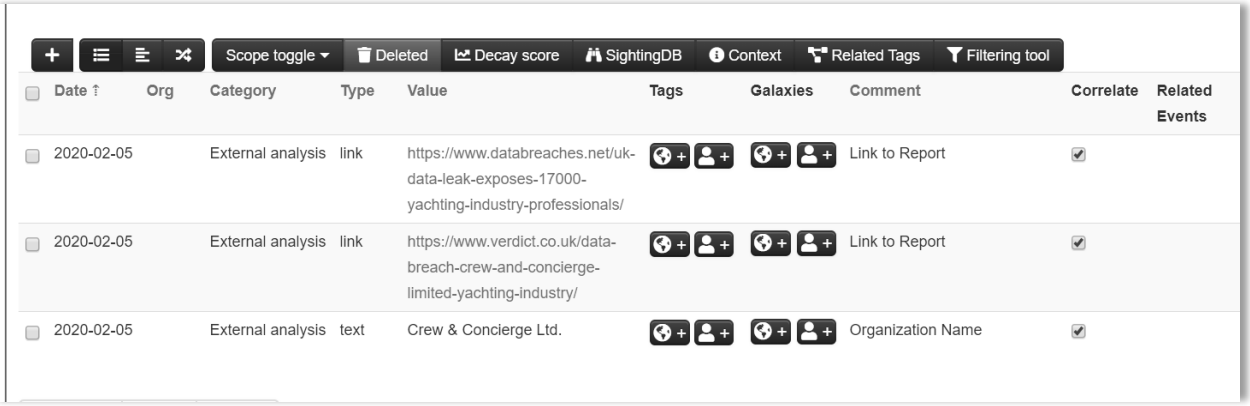
As described in Chapter 1, a data breach artifact conveys information regarding the violation or breakdown of the CIA triangle. For this example, we will be using the data breach artifact *UK: Data leak exposes 17,000 yachting industry professionals*. This artifact was chosen for the data breach category because of the language used in the title, which conveys an occurrence of a data breach event. Navigating to the artifact's web page, we see that there is only general information about the event. Upon further examination of the artifact, we see that there is a reference to the source artifact obtained by the primary artifact's author. Therefore, to gain new insight, we navigate to the source site. We can see that the original source is from a website titled *Verdict.* However, before proceeding to this new website, we have an opportunity to begin annotation and have enough information to start the event using a template. *Using the information from Chapter 1, we select the template that begins a Data Breach Non-US event.* We also have enough information to start the annotation process that will eventually lead to the tagging of the event. Here we can see that the organization is *Crew and Concierge Limited*, and one of the sources of information is *Databreaches.net*. However, the analyst will also notice there are missing pieces of information, such as the BLS industry code for Crew and Concierge Limited.



For the sake of simplicity, we will use the application *Notepad* provided as an inherent feature of Microsoft Windows as our note-taking tool. The level of detail used during any annotation activity will vary by the analyst. This book does not suggest nor recommend a system of annotation and leaves that to individual analyst familiarity and creativity. Below are some examples of the notes taken during this initial engagement with the artifact(s).

The following images show how the event will look after the initial information has been entered in using the appropriate template. The first image is of the general event information. Here we can see the title of the event, the date for which the event was created, and the title of the event using the first artifact's title in the text. Additionally, we can see that both the first and second sources of information are referenced by their corresponding web link, as well as the name of the organization that corresponds with the data breach. Next we will look to determine the context of the event and record the relevant text.





In comparison, the first source had little information regarding the data breach as opposed to the amount of the information supplied by the original source. This lack of information is not an error, nor a negative aspect of the first source ability to report the artifact. Instead, this is quite typical when working with OSINT artifacts. Often reporting sources will summarize the information given by the original source. Truthfully, it is better to find the data than to criticize the reporting of the data. For example, due to the way for which the information is gathered from the internet, if it were not for the posting by the reporting source, the original source information may have never been found. In any case, let's now begin to deconstruct the artifact to get at the information required to complete the event. For this, we will use the more verbose artifact, the original source artifact.

The first thing we will look at is the title of the artifact. The title of the original artifact is *Exclusive: Data breach exposes 17,000 yachting industry professionals.* From this title, we can annotate a couple of things. The first thing we notice is the relative number of affected individuals (17,000). Secondly, we can see it was not customers of the organization, but the employees who were directly impacted. Moving further into the artifact, we find a more accurate number of victims represented in the data. We can also gain a general sense of the organization's type. In this example, the organization appears to be within the human resources

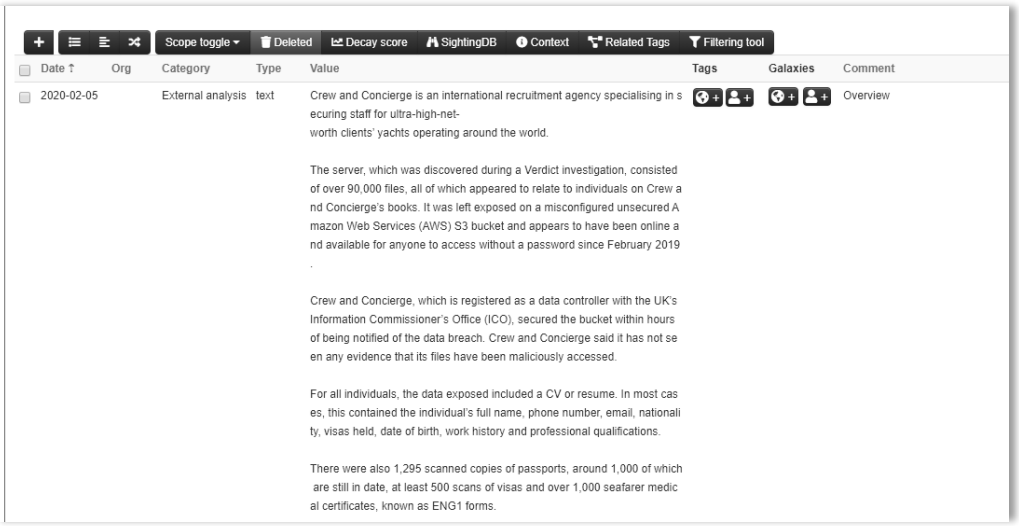
industry, more specifically, a recruiting company that specializes in staffing for yachting operations globally.

> A data breach at UK-based Crew and Concierge Limited has exposed the personal data of 17,379 people of 50 different nationalities working in the yachting industry.
>
> Crew and Concierge is an international recruitment agency specialising in securing staff for ultra-high-net-worth clients' yachts operating around the world.

Continuing to work through the artifact, we find the circumstance for the data breach. In this case, it was a misconfigured/unsecured cloud-based server that was discovered online.

> The server, which was discovered during a *Verdict* investigation, consisted of over 90,000 files, all of which appeared to relate to individuals on Crew and Concierge's books. It was left exposed on a misconfigured unsecured Amazon Web Services (AWS) S3 bucket and appears to have been online and available for anyone to access without a password since February 2019.

Lastly, towards the end of the artifact, we have information regarding the type of information exposed. In this case, it was the resume and passport information which contained but is not limited to full name, email, nationality, date of birth, and work history. All of this information will be used later to add an actual data breach severity score to the data breach event. At this point in the process, we add all of the text from the artifact to the event. To do this, we would add a single attribute to the event setting the flag for *Category* to *External Analysis*, the *Type* to *text*, the *Distribution* to *Inherit event*, and the *Contextual Comment* would be *Overview*. The following image shows how the textual information from the artifact appears as after adding it to the event.



The event has been annotated, recorded, and is now ready for the addition of the galaxy tags. Of all the procedures performed during the recording process, the attachment of galaxy tags is probably the hardest to accomplish. Due to the nature in which the information is reported, there are opportunities for competing or congruent interpretations of which tags should be attached to the event; and, to what extent.

Sincerely, this is a new area of research and application; therefore, the attachment of galaxy tags is best-effort. It is recommended to any analyst or organization to go back through the textual content of an artifact and discuss the application of galaxy tags and the possible shortcomings or improvements that can be made within a cycle of continuous refinement.

Additionally, it is recommended to send these outputs of internal discussion to the organizations that publish the frameworks for which the galaxy tags are built. The collaboration includes and informs these organizations so that they, too, can perform their cycle of continuous improvement. With this in mind, each of the galaxy tags attached to the example event is done so with some form of analyst interpretive bias and good intent.

Due to the large number of possible galaxy tags that can be selected the easiest way to stay organized is to work down the CMCF by section and apply the relevant galaxy tags. The following order of steps would be appropriate for any artifact:

1. General Artifact
2. Author/Source
3. Block-H
4. Combined Operational Framework
5. MITRE Pre ATK
6. MITRE ATK
7. SecTool
8. CWE

Informational drawbacks as a section have been left out of that order. The reason for this is that informational drawbacks are discovered during the process of attaching galaxy tags to an event or during the closing section of event recording. As we walk through each of the CMCF categories, you will see how the informational drawbacks are applied.

Section 2: Adding General Artifact Tags to the Event

The first category to examine for galaxy tag attachment is the General Artifact category. This category contains the sub-categories of Artifact Type, General Data Breach Index (GDBI), Government Entity Classification Index (GECI), U.S. State Index, ENISA DB SE Rating, and the Bureau of Labor Statistics Index. From this category the following galaxy tags can be attached to the event:

Data Breach (Non-US)
GDBI: Misconfiguration
NIACS 54 Professional, Scientific, and Technical Services

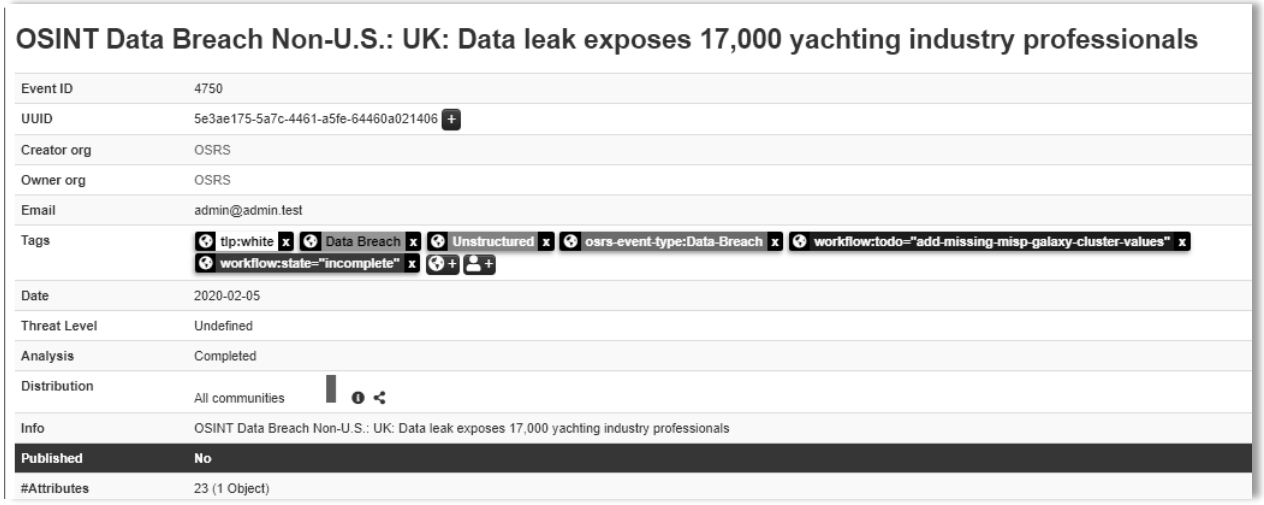
The GECI cluster does not apply as the organization involved in the data breach event is not a government entity. The U.S. State Index is not relevant either as the data breach event did not occur within the sovereign boundaries of the United States. Lastly, the ENISA DB SE Rating does apply, but that is covered in a different section of this chapter regarding the scoring of a data breach event.

Section 3: Adding Author/Source Tags to the Event

Under the Author/Source category of the CMCF, we can attach the following galaxy tag to the event:

OSINT Source: Databreaches.net

However, during this procedure, we notice that we do not have any tags that relate to the source *Verdict*. Therefore, the analyst would include in their annotation of the event that there needs to be a galaxy tag created to reflect this source. Additionally, the analyst would use the built-in workflow taxonomy tags provided by MISP and tag the event as *incomplete* as well as *having a missing galaxy cluster*.



Section 4: Adding Block-H Tags to the Event

Examining the artifact from the perspective of the Block-H category generates some ambiguity. The sub-categories under the parent category of Block-H include Risk Management Program, Information Protection Policy, Information Protection Organization, Human Resources, Asset Management, Access Control, Cryptography, Physical/Environmental Protection, Operations, Communications, System Acquisition/Development, Supply Chain, Incident Management, Business Continuity, Compliance, and Privacy. To work through this parent category efficiently and accurately, we will need to consider each sub-category from the perspective of the violation or breakdown of the CIA triangle as it relates to Block-H. In short, we will repeatedly ask ourselves if any of the individual items within each sub-category were somehow violated or broken until we have gone through the entirety Block-H. The following table will help us complete this series of questions.

Block-H Sub-Category	Were any of the sub-categories violated, or did any of the sub-categories relate to a breakdown of the CIA triangle?	Are there instances or opportunities for informational drawbacks?
Risk Management Program	It was unclear in the artifact what if any risk management constructs were followed by the organization.	The artifact does not directly state any information regarding risk management. However, there was a sense that the organization did have some sort of risk management built into their organizational model.
Information Protection Policy	It was unclear in the artifact what, or to what extent the organization applied information security policy.	The artifact does not directly state any information regarding information security policy. However, there was a sense that the organization did have some sort of information security policy built into their organizational model.
Information Protection Organization	The artifact alludes to the importance of information protection but does not clearly lay out an information protection organization within the parent body.	The artifact does not directly state any information regarding information protection organization. However, there was a sense that the organization did have some sort of information protection organization built into their organizational model.
Human Resources	The article does state that the affected part of the organization was a server owned by the	

	organization. Securing this publicly-facing server should have been covered within some form of security awareness training.	
Asset Management	The asset was a server that was owned by the affected organization. However, this asset clearly did not comply with the proper handling of assets that contain this type of information.	
Access Control	This server did not require a password to access.	
Cryptography	There were no cryptographic controls applied to the server.	
Physical/Environmental Protection	Another service hosted their server.	It is unclear as to what physical protections the cloud service provider may provide.
Operations	There was no need to use a password to log into the server. The lack of password requirements is considered a technical, as well as a procedural vulnerability.	
Communications	A cloud service provider hosts the server.	It is unclear as to what controls are used by the cloud service provider to prevent unwanted or unauthorized communications.
System Acquisition/Development	The server was hosted and publicly available with no password required for logging into the server.	
Supply Chain	The system was owned by the organization and was not provided by a third party.	
Incident Management	The organization did respond to the incident, both publicly and internally.	
Business Continuity	This event does not affect the organization's ability to continue its operations.	
Compliance	Various records were containing a variety of information types accessed during the data breach event.	
Privacy	Privacy protection reporting procedures were followed by the organization both publicly and privately	

The following is a list of Block-H galaxy tags that can be attached to the event:

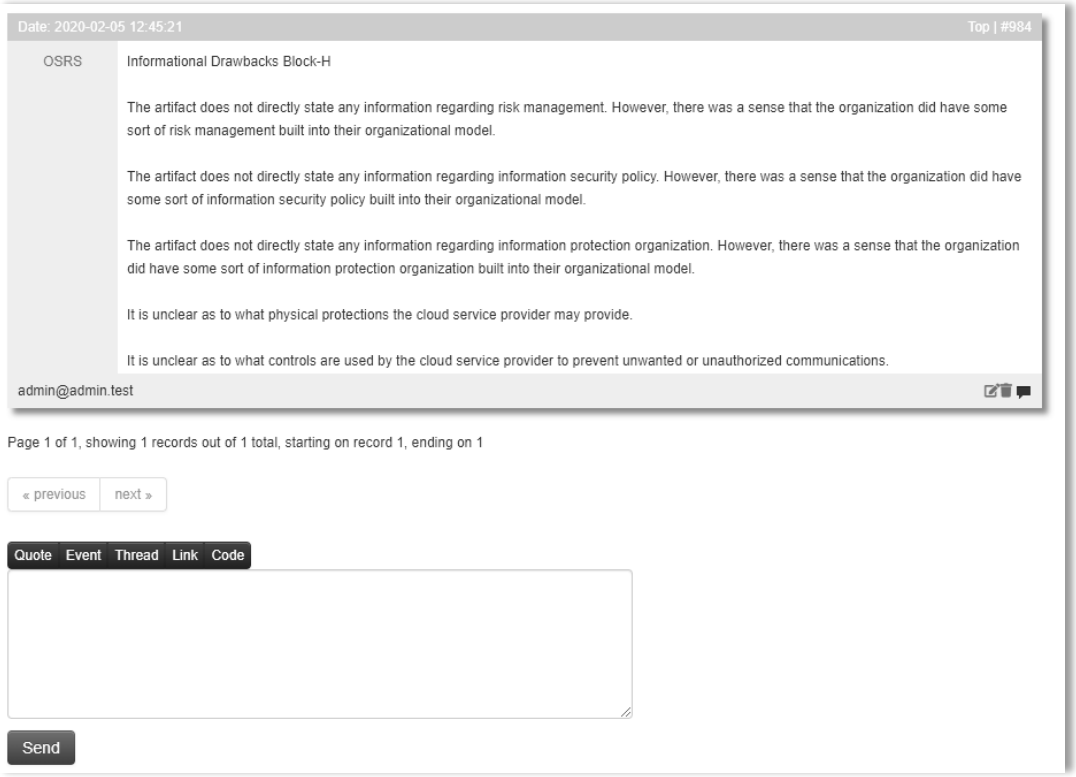
03.c Management Responsibilities
03.d Information Security Awareness, Education, and Training
04.g Handling of Assets
05.e Management of Privileged Access Rights
05.i Use of Secret Authentication Information
05.j Information Access Restriction
05.k Secure Log-On Procedures
06.a Use of Cryptographic Controls
06.b Key Management
08.l Management of Technical Vulnerabilities
10.b Securing Application Services on Public Networks
10.i Secure System Engineering Principles
10.l System Security Testing
10.m System Acceptance
14.c Protection of Records
14.d Protection of Personally Identifiable Information

As we can see, this organization, based on the attached galaxy tags, failed to handle and secure the organizational asset. This behavior resulted in several violations and breakdowns of the CIA

triangle. There were also opportunities to recognize informational gaps in the information conveyed by the artifact. From the far-right column, we can see that there were several instances of hinting at certain areas. Still, not enough language used to convey a direct correlation between the artifact and the appropriate Block-H galaxy tag. Therefore, we would attach the following informational drawbacks to the event as well as annotate their notes within the notes section of the event.

Informational Drawback: Block-H Artifact Vague

This part of the chapter is an excellent area to discuss the addition of notes to an event. MISP has a region located at the bottom of the main event page that allows the analyst to add notes to an event. These notes can be handy for documenting informational drawbacks, missing tags, workflow needs, or any type of annotation that an analyst wishes to attach to the event. After adding the Block-H tags, we also add the additional notes to the event using this built-in MISP capability. The following image shows the output of that procedure.



Section 5: Adding Combined Operational Framework Tags to the Event

Moving on, we come to the category within the CMCF called the Combined Operational Framework. Cybersecurity frameworks vary in language based on the type of audience that will apply the framework to their specific cybersecurity role within the environment. For example, the MITRE ATK framework is used to describe and classify events at a very technical level and

mainly focuses on describing malware-based activity. The Combined Operational Framework is a blend of two structures to provide a middle-management level of descriptive information to the event.

The same methodology used for working through Block-H can be used to work through the sub-categories of the Combined Operational Framework. The following sections require a walkthrough.

- Environmental Threat
- Failure Threat
- Organizational Threat
- Existential Threat
- Human Threat
- Nefarious Threat

A table can also be used to help organize the examination of the artifact from the perspective of the Combined Operational Framework and the violation or breakdown of the CIA triangle.

Combined Operational Framework Sub-Category	Were any of the sub-categories violated, or did any of the sub-categories relate to a breakdown of the CIA triangle?	Are there instances or opportunities for informational drawbacks?
Environmental Threat	The data breach was not the result of an environmental threat	
Failure Threat	The data breach was not the result of a failure type threat	
Organizational Threat	The data breach was not the result of an organizational threat	
Existential Threat	The data breach was not the result of an existential threat	
Human Threat	The data breach was the result of human behavior	
Nefarious Threat	The artifact does convey the data breach event by a nefarious threat	

From the completed table, the analyst can attach the following tags to the event:

ENISA Erroneous Use or Administration of Devices and Systems
ENISA Inadequate Design and Planning or Improper Adaptation
LUH4 Mishandling of Passwords
LUM15 System Configuration Errors
LUM7 Improperly Designing Information Systems
ENISA Compromising Confidential Information (Data Breaches)
LIN49 Unauthorized Access

For this section, there were little informational drawbacks, and the information provided within the artifact contained enough clarity to attach Combined Operational Framework galaxy tags to the event accurately.

Section 6: Adding MITRE Pre ATK Tags to the Event

The MITRE Pre ATK framework is used for describing pre-attack adversarial actives. In laymen's terms, these descriptors tell the story of how the adversary prepared themselves to attack their intended target. Again, we will apply the same method of question and answer to work through the sub-categories of the MITRE Pre ATK framework contained within the CMCF. The following table can be used to assist the analyst in this process. The most noticeable difference between the MITRE Pre ATK table and the previous tables for Block-H and the Combined Operational Framework is that we are merely trying to identify if any relevant information maps to a MITRE Pre ATK category.

MITRE Pre ATK Framework Sub-Category	Were any of the sub-categories of the MITRE Pre ATK Framework sufficiently described within the artifact.	What type of information drawbacks may exist for this sub-category?
Priority Definition Planning	The artifact eludes to the fact that the original source discovered the server online. It does not describe how the attack was planned.	Information is missing or vague.
Priority Definition Direction	The artifact eludes to the fact that the original source did have a set of technology for discovering and exploited the particular technical weakness but does not describe the particulars of their toolset.	Information is missing or vague.
Target Selection	The organization had possible targets in mind. How the source organization determined these targets is unclear.	Information is missing or vague.
Technical Information Gathering	The artifact vaguely mentions the fact that the source organization could derive technical information about a target, but the exact procedure or toolset was not mentioned. However, to find the target, the source organization had to have conducted active scanning as the server was a publicly available target. The source organization also discovered that there was no password required to log on to the server. The source organization also discovered that a 3rd party hosting service hosted the server.	
People Information Gathering	There was no mention as to the need for people information gathering to conduct the attack.	Not relevant to the data breach event.
Organizational Information Gathering	The artifact vaguely mentions the fact that the source organization had an idea of the type of organization they were targeting.	It was covered in technical information gathering.
Technical Weakness Identification	The source organization was able to determine that a password was not required to logon to the system.	
People Weakness Identification	The artifact did not convey the need to identify people's weaknesses within the organization.	Not relevant to the data breach event.
Organizational Weakness Gathering	The artifact did not convey the need to identify organizational weaknesses within the organization.	
Adversary OpSec	The artifact conveys little about adversary OpSec. Nor does it allude to the fact that any OpSec was used at all.	Information missing.
Establish/Maintain Infrastructure	The artifact conveys a sense that the original source did have some sort of established infrastructure to conduct the attack. However, to perform this type of technical attack, the original source did have to procure the required equipment and software. These tools would also need some	

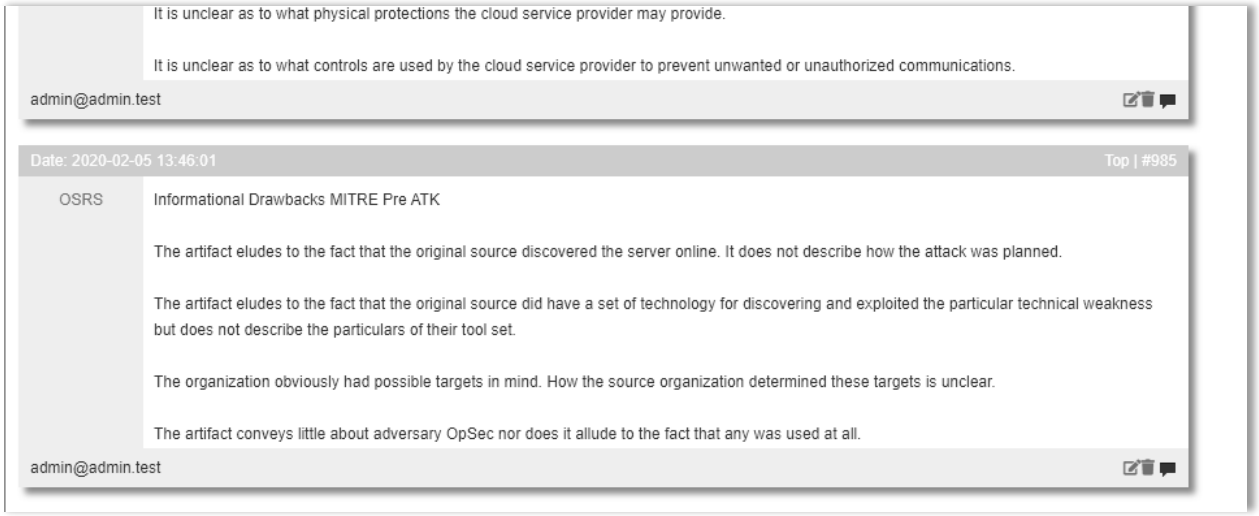
	sort of installation and configuration to work correctly.	
Persona Development	There was no need to develop a persona to conduct the attack.	Not relevant to the data breach event.
Build Capabilities	There was no requirement to build systems mentioned within the artifact. Nor did the artifact allude to the fact that this was necessary to conduct the attack.	Not relevant to the data breach event.
Test Capabilities	There was no requirement to build systems mentioned within the artifact. Nor did the artifact allude to the fact that this was necessary to conduct the attack.	Not relevant to the data breach event.
Stage Capabilities	There was no requirement to build systems mentioned within the artifact. Nor did the artifact allude to the fact that this was necessary to conduct the attack.	Not relevant to the data breach event.

Based on the results of the table entries, the following tags can be attached to the event:

Technical Information Gathering: T1254 Conduct Active Scanning
Technical Information Gathering: T1255 Discover Target Logon and (or) Email Address Format
Technical Information Gathering: T1260 Determine 3rd Party Infrastructure Services
Technical Weakness Identification: T1288 Analyze Architecture and Configuration Posture
Establish & Maintain Infrastructure: T1335 Procure Required Equipment and Software
Establish & Maintain Infrastructure: T1336 Install and Configure Hardware, Network, and Systems

This section also allows for further annotation and attachment of informational drawbacks. To complete this procedure, we would follow the same method used for annotating and attaching informational drawbacks found while attaching Block-H galaxy tags. Based on the results of the entries made on the table, the following informational drawbacks can be attached to the event. Additionally, an image of the attached annotations can be found below the list of informational drawback tags attached to the event.

Informational Drawback: MITRE Pre ATK Artifact Vague (Priority Definition Planning)
Informational Drawback: MITRE Pre ATK Artifact Vague (Priority Definition Direction)
Informational Drawback: MITRE Pre ATK Artifact Vague (Target Selection)
Informational Drawback: MITRE Pre ATK Information Absent (Adversary OpSec)



Section 7: Adding MITRE ATK Tags to the Event

As previously mentioned, the MITRE ATK is the second piece to the entire MITRE framework used to describe the attack. The difference between MITRE Pre ATK and MITRE ATK is the subject. The MITRE Pre ATK is used to describe the preparation an adversary may conduct before launching an attack, whereas MITRE ATK is the description of the actual attack. In total, there are twelve sub-categories to the MITRE ATK category contained within the CMCF. These sub-categories are further broken down into individual entries by the system Windows, Linux, AWS, GCP, Azure, and SaaS.

MITRE ATK Framework Sub-Category	Were any of the sub-categories of the MITRE ATK Framework sufficiently described within the artifact.	What type of information drawbacks may exist for this sub-category?
Initial Access	The target was a publicly facing AWS bucket.	
Execution	Execution information absent to the extent that it does not describe if the AWS bucket was a windows-based bucket or Linux-based, only that it was available.	Information absent
Persistence		
Privilege Escalation	Valid accounts are considered usable because no password was required, thus automatically making any account valid.	
Defense Evasion	Defenses were evaded by using valid account via the non-existence of a required password.	
Credential Access	No relevant category entry for this AWS type of event.	No relevant category
Discovery	The bucket was discovered remotely.	
Lateral Movement	There was no clear evidence of lateral movement reported within the artifact, simply that the source exploited a single bucket. It is not to say that they did not attempt to move laterally into other systems owned by the target organization.	Information vague
Collection	The data was hosted on a cloud storage object and retrieved from that object.	
Command and Control	No relevant category for this particular MITRE ATK category.	No relevant category.

Exfiltration	The data was not taken only observed; thus, Exfiltration does not apply to this event.	
Impact	No relevant category as resources were not used by the source organization who initiated the attack.	No relevant category.

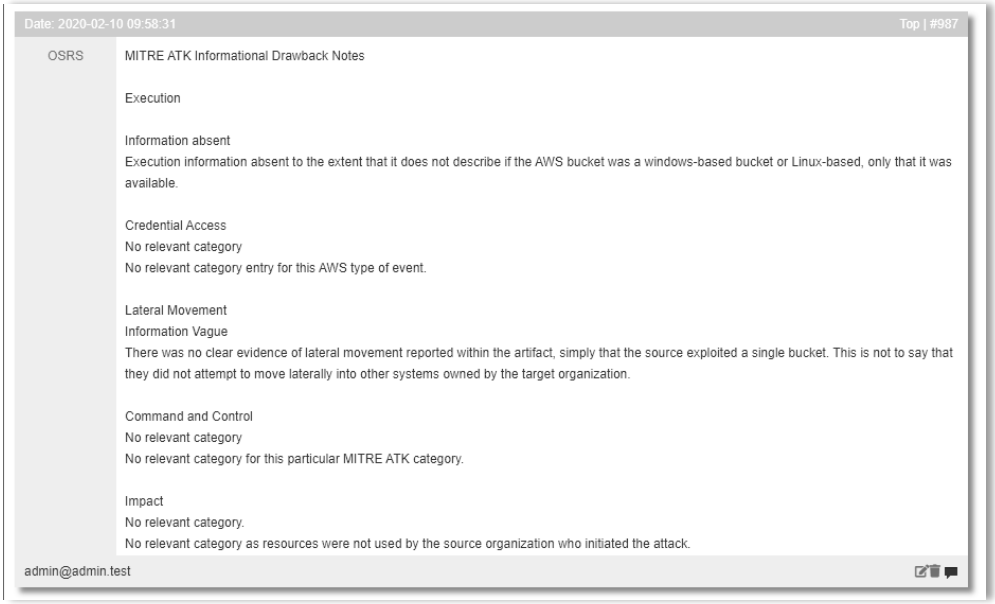
From the table the following MITRE ATK entries can be attached to the event:

Initial Access: AWS T1190 Exploit Public Facing Application
Privilege Escalation: AWS T1078 Valid Accounts
Defense Evasion: AWS T1078 Valid Accounts
Discovery: AWS T1018 Remote System Discovery
Collection: AWS T1530 Data from Cloud Storage Object

From the perspective of the information gathered within the table, the following informational drawbacks can be attached to the event:

MITRE ATK Artifact Vague (Lateral Movement)
MITRE ATK Information Absent (Execution)
MITRE ATK No Relevant Category (Credential Access)
MITRE ATK No Relevant Category (Command and Control)
MITRE ATK No Relevant Category (Impact)

After all relevant MITRE ATK galaxy tags have been attached to the event, we then apply the notes regarding any informational drawbacks. This process is the same for the information drawbacks reported in the Block-H and MITRE Pre ATK informational drawback notes. The following image shows the results of adding the MITRE ATK informational drawbacks to the event.



Section 8: Adding MalTech Tags to the Event

This data breach was not the result of malware or threat actor. Therefore there is neither an opportunity to tag the event with informational drawbacks; or specific MalTech nomenclatures. Thus, the analyst can skip this portion of the procedure. However, if the data breach causality were attributable to something like a particular family of ransomware we would have the opportunity to attach the appropriate MalTech cluster or discover opportunities to connect informational drawbacks. For example, a data breach artifact states that the target received ransomware, but the name of the ransomware family was not identified in the artifact. We could then attach a MalTech informational drawback saying that the information regarding the ransomware nomenclature was absent from the artifact.

Section 9: Adding SecTool Tags to the Event

The event is not reflective of this category. However, if, by chance, a SecTool like MimiKatz were to be identified in the artifact, we would add in this information. Truthfully, it is uncommon to have this level of detail in any data breach report.

Section 10: Adding CWE Tags to the Event

A data breach typically is a result of some kind of vulnerability. This vulnerability may sometimes be technical. In the case of this example, the weakness is the lack of requiring a password to log in. Therefore the analyst can apply the following CWE galaxy tag:

CWE-521 Weak Password Requirements

This CWE is sufficient enough to describe the weakness that was exploited to gain access to the data. Like SecTool information, it is uncommon to see this type of data contained within a data breach report. As a result, many data breaches will have CWE informational drawbacks attached rather than a specific CWE galaxy tag.

Scoring a Data Breach

The recording of a data breach also includes the assignment of a data breach severity score. This process is based upon the initial recommendation for data breach severity scoring produced by cybersecurity researchers affiliated with ENISA, a European cybersecurity council. This process is not an easy process to work through, as there are several individual scores used to calculate the overall severity rating score and ranking. The following table represents the workflow and scoring per section of the ENISA SE DB Rating section within the CMCF.

ENISA SE DB Rating Category	Analyst Notation	Category Score
Data Processing Context: Simple Data	Base score = 1 Adjusted by +1 because the information exposed contained information that could be used to profile the individual as well as make assumptions about the individual's social/financial status.	2
Data Processing Context: Behavioral Data	Information absent and not exposed.	0
Data Processing Context: Financial Data	Information absent and not exposed.	0
Data Processing Context: Sensitive Data	Base score: 4 Health information via seafarer medical certificates exposed. Decreased by 1 because general assumptions can be made based on this type of sensitive information.	3
Data Processing Context: Total Data Processing Context Score		5
Ease of Identification: Full Name	Information about the commonality of the exposed information not reported.	0 (Absent/Unknown)
Ease of Identification: ID Card, Passport, Social Security Number	Full passport copies exposed.	+.75 (Significant)
Ease of Identification: Telephone Number or Home Address	Phone number information exposed. Unknown if it is publicly available.	+.25 (Negligible)
Ease of Identification: Email Address	Email address exposed but unknown if it directly ties the individual to other accounts or information.	+.25 (Negligible)
Ease of Identification: Picture	Information absent and not exposed.	0 (Absent/Unknown)
Ease of Identification: Coding, Aliases, or Initials	Information absent and not exposed.	0 (Absent/Unknown)
Ease of Identification: Total Modifier		1.25
Circumstances of Breach: Loss of Confidentiality	Confidentiality lost due to unsecured publicly available cloud resource.	+.5
Circumstances of Breach: Loss of Integrity	Data was not altered. No loss of integrity.	0
Circumstances of Breach: Loss of Availability	Data was never unavailable.	0
Circumstances of Breach: Malicious Intent	No malicious intent related to the exposure	0
Circumstances of Breach: Total Modifier		.5
Severity Rating		6.75
Severity Ranking		Very High

Now that we have captured the data, we need to complete the equation and give the event a data breach severity score. The equation for this is *SE = (DPD x EI) + CB*, or rather the Severity Rating equals the Data Processing Context (DPD) times the Ease of Identification (EI) modifier plus the Circumstance of Breach (CB). From the results found in the table, our DPD total score equals 5. The EI modifier is 1.25. Lastly, the CB value is .5. Plugging these numbers into the equation, we get the following result:

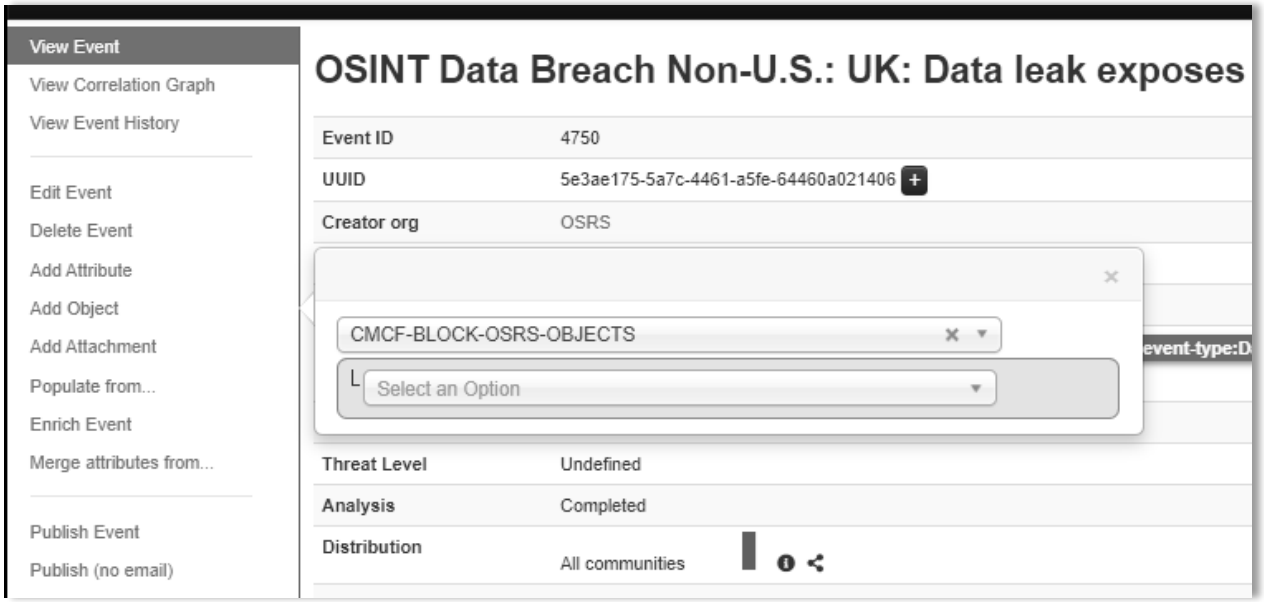
$$SE = (5 * 1.25) + .5$$

$$SE = 6.75$$

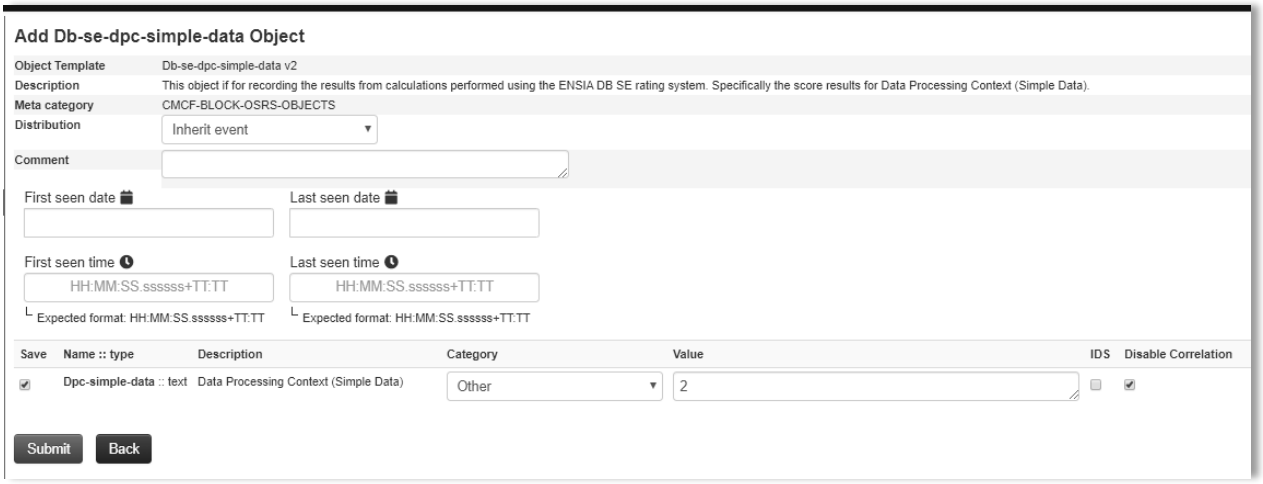
From the result, we would say that the severity rating ranking falls within the parameters of *Very High* based on the SE value of 6.75. Once we have these final values, we attach the values to the event with the appropriate galaxy tag. However, there is a very critical difference in the way for

which these tags are added. As opposed to using the general galaxy tag section of the event as seen previously, we will tag the actual attribute. We do this for future data mining activities using Splunk, and we will see how that works later chapters. For now, we will simply add the information and tags to the event. Unlike adding the data using a template, we must use a different function in MISP called *Objects.* The concept of objects is not within the scope of this book. The reasoning for using objects versus templates is that objects allow for the inclusion of duplicate data. For example, our series of scores contains several repeated *0* values. If we were to use a temple, MISP would reject the initial submission stating that the attribute value already exists, or it will not record the values at all. Objects do check for this type of value redundancy.

The OSRS maintains a set of MISPs object called the *CMCF-BLOCK-OSRS-OBJECTS*, which can is found under the *CMCF-BLOCK-OSRS-OBJECTS* category within the MISP object selection tool. The object selection tool is located to the left of the event's main page.



There are several objects under our parent directory. The analyst will simply select the object that relates to a section on the scoring table and enter the data. Once the appropriate object has been chosen, we are directed to a separate web page that allows for the actual data entry. The following image shows what this page looks like an analyst, as well as where the values will be placed within the object. As we can see in the picture, we only need to be concerned with two columns. The first column, titled *Name::type,* located to the far left, identifies the exact section from the previous scoring table. The second column of interest, titled *Value,* is the location that the analyst will enter the data. All other parts remain untouched.



Like adding information to an event via a template, we are allowed to review and modify the data before adding it to the event. Severity Ranking is somewhat different. Here we will type in the textual value of rank versus a numerical result. The ranking system is as follows:

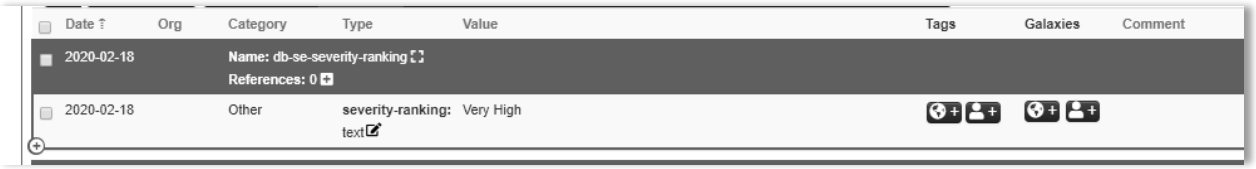
$$SE < 2 \text{ Low}$$

$$2 \leq SE < 3 \text{ Medium}$$

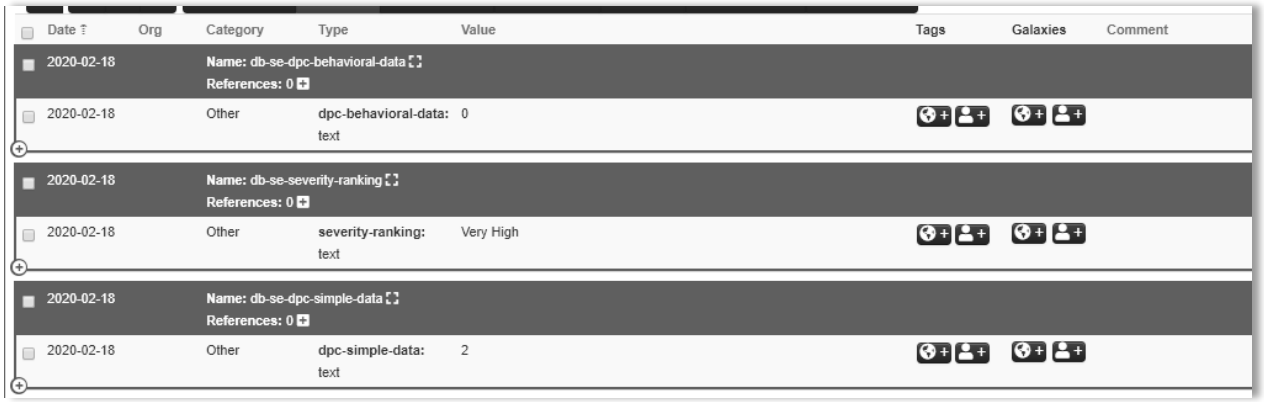
$$3 \leq SE < 4 \text{ High}$$

$$4 \leq SE \text{ Very High}$$

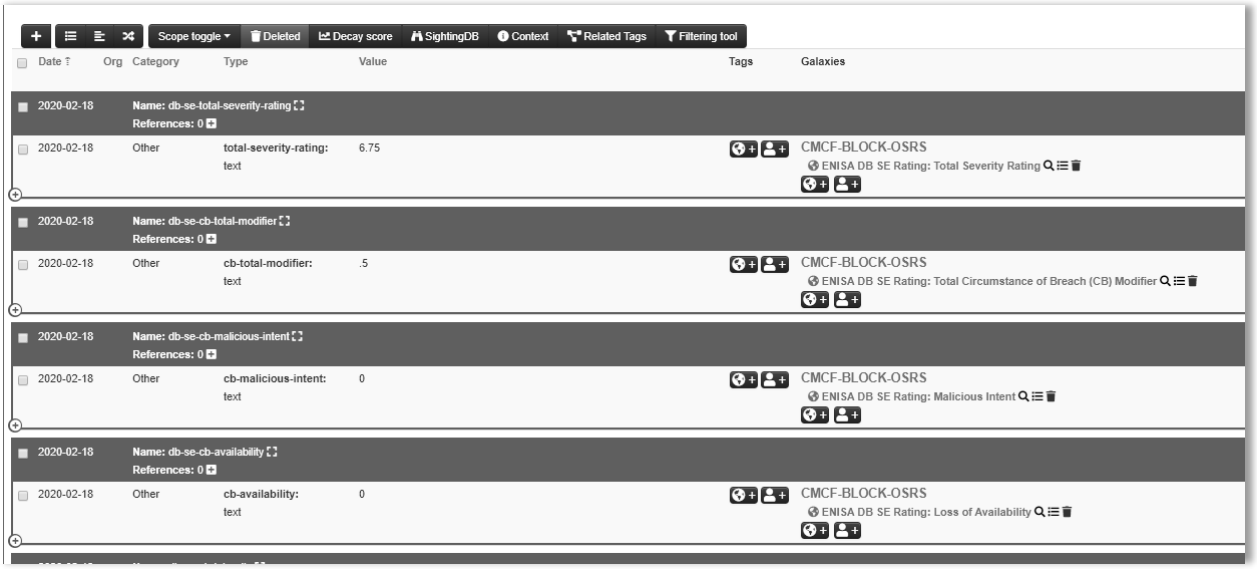
Based on the SE value, our severity ranking is determined to be *Very High*. The data entry is reflected in the image below.



Once confident in the accuracy of the data, we will click the button labeled *Create new object* to have the information added to the event.



Notice how the object allowed for duplicate attributes to be added to the event. The last piece to the scoring of the event is to add the appropriate tags to the individual scores. These tags are found within the *General Artifact* section of the CMCF under the sub-category of *ENISA DB SE Rating.*



With all the appropriate galaxy tags added to the attributes, we are now ready to publish the event and end the artifact recording process. To do this, we simply click the *Publish Event* link located to the left-hand side of the main event page. However, remember that there was a missing tag within the CMCF. That was the tag related to an organization that had not been previously recorded as an entity within MISP. There is no right way to determine when an analyst is required to go back and complete this portion of the event. Each analyst has to manage the completeness of artifact recording within the constraints of their own time management and organizational goals.

Chapter Summary

In this chapter, we learned the process and procedures for recording a data breach artifact. This methodology is labor-intensive, but like all the artifact recording processes and methods presented in this book, they set up the information for future use. Each analyst will approach the recording of artifacts differently. Some may prefer to initially score the artifact to remove that from the set of procedures. Conversely, others might want to simply enter in the textual information and then circle back to the scoring of the artifact. There is no one exact way to do it. What is most important is the completeness and time the analyst spends to ensure that all of the data is entered accurately and with enough critical thinking behind the activity to secure good efficacy and effectiveness as portrayed through the data.

Chapter Acronyms

ATK: MITRE ATT&CK Framework

AWS: Amazon Web Services

BLS: U.S. Bureau of Labor Statistics

CB: Circumstance of Breach

CIA: Confidentiality, Availability, and Integrity

CMCF: Comprehensive Modular Cybersecurity Framework

CWE: Common Weakness Enumeration

DPD: Data Processing Context

EI: Ease of Identification

ENISA: The European Union Agency for Cybersecurity

GCP: Google Cloud Platform

GDBI: General Data Breach Index

GECI: Government Entity Classification Index

HITRUST: Health Information Trust Alliance

MISP: Malware Information Sharing Platform

NAICS: The North American Industry Classification System

OPSEC: Operational Security

OSINT: Open Source Intelligence

SAAS: Software as a Service

SE: Severity Rating

Chapter References

Amazon Web Services (AWS)—Cloud Computing Services. (n.d.). Retrieved February 18, 2020, from https://aws.amazon.com/

Cloud Computing Services | Google Cloud. (n.d.). Retrieved February 18, 2020, from https://cloud.google.com/

Common Weakness Enumeration. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Weakness_Enumeration&oldid=918897150

Data breach exposes 17,000 yachting industry professionals. (2020, February 4). *Verdict*. https://www.verdict.co.uk/data-breach-crew-and-concierge-limited-yachting-industry/

ENISA. (n.d.). Retrieved February 4, 2020, from https://www.enisa.europa.eu/

Galan Manso, C., Górniak, S., & European Network and Information Security Agency. (2013). *Recommendations for a methodology of the assessment of severity of personal data breaches.* ENISA. http://bookshop.europa.eu/uri?target=EUB:NOTICE:TP0413118:EN:HTML

HITRUST. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=HITRUST&oldid=938925094

Home | U.S. Bureau of Labor Statistics. (n.d.). Retrieved February 4, 2020, from https://www.bls.gov/

Information security. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Information_security&oldid=939133610

MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing (formely known as Malware Information Sharing Platform). (n.d.). Retrieved February 4, 2020, from https://www.misp-project.org/

North American Industry Classification System. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=North_American_Industry_Classification_System&oldid=932760038

Open-source intelligence. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open-source_intelligence&oldid=937269341

open-source-rs. (2020). *Open-source-rs/Comprehensive-Modular-Cybersecurity-Framework-CMCF*. https://github.com/open-source-rs/Comprehensive-Modular-Cybersecurity-Framework-CMCF (Original work published 2019)

Operations security. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Operations_security&oldid=922796009

Software as a service. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Software_as_a_service&oldid=941066057

Union, P. O. of the E. (2014, February 3). *Recommendations for a methodology of the assessment of severity of personal data breaches :* [Website]. https://op.europa.eu:443/en/publication-detail/-/publication/dd745e70-efb8-4329-9b78-79020ec69da5/language-en

https://www.linkedin.com/company/threathunting
https://www.twitter.com/threathunting_

Chapter 3: Methodology for Recording MalTech OSINT Artifacts

Chapter 3 Introduction

In this chapter, we will look at how an analyst might record a MalTech OSINT artifact using the OSRS methodology for recording intelligence. MalTech, unlike the Data Breach category, is the most diverse category in terms of the type of information that can be collected. Additionally, this category, unlike any of the other categories shown in the hierarchy of categories from Chapter 1, has the highest propensity to comingle child categories within a single artifact. A threat actor may deploy a botnet, or a campaign leverages ransomware delivered by a botnet to extort financial resources from its victims.

Regardless of how the information may comingle categorically, this chapter's goal is to provide an analyst with enough example information that the analyst can apply the methodology to any item considered MalTech categorically. We will also use the same method of structured inspection of an artifact using annotation tables, as seen in Chapter 2. Lastly, through the examples in this chapter, we will see how analyst experience influences the efficacy of the recording process and procedure, as well as how analyst bias plays a role in overall effectiveness as well.

Recording the MalTech Artifact

Section 1: Artifact Selection, Event Creation, and Adding General Artifact Tags to the Event

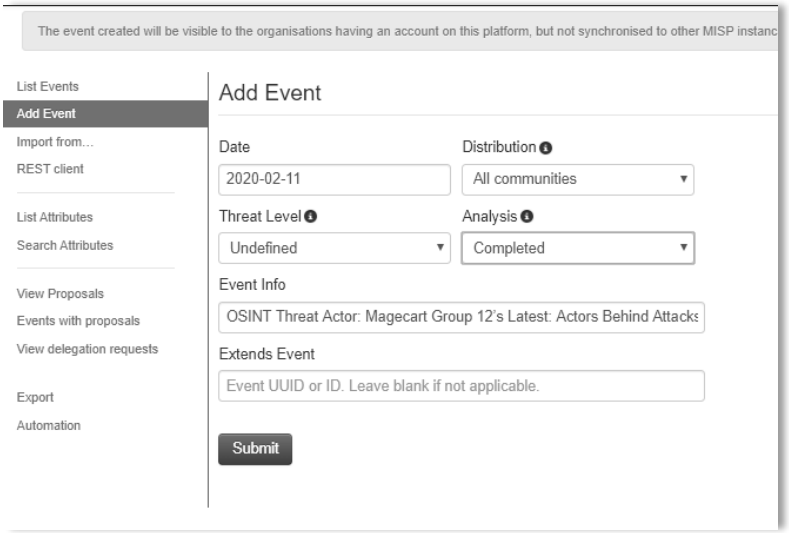
MalTech encompasses a set of sub-categories used for classifying OSINT artifacts that describe threat actors, backdoors, botnets, campaigns, ransomware, and tools that exploit cyber systems. Just like the recording of data breach artifacts, an analyst starts by identifying an artifact that aligns to one of the sub-categories. As previously mentioned in previous chapters, there is no singularly good way to find OSINT artifacts. Each analyst will have a varying degree of success and failure in terms of OSINT collection success. For this example, we will be looking at an OSINT artifact titled *Magecart Group 12's Latest: Actors Behind Attacks on Olympics Ticket Re-sellers Deftly Swapped Domains to Continue Campaign*.

Recording MalTech artifacts are very similar in process to the recording of data breach artifacts. First, we will begin by creating the initial event. Again, it is important to use the correct title schema when creating the event. The following schema and examples previously laid out in Chapter 1 are used for titling MalTech artifacts:

5) MalTech
    a. Actor, Campaign, or Both
        i. Actor
            1. Syntax: OSINT Threat Actor: <title from artifact>
            2. Example: OSINT Threat Actor: Mr. Fox Strikes Again
        ii. Campaign
            1. Syntax: OSINT Campaign: <title from artifact >
            2. Example: OSINT Campaign: Orange Peel Botnet Campaign
        iii. Both
            1. Syntax: OSINT Actor/Campaign: <title from artifact>
            2. Example: OSINT Actor/Campaign: Mr. Fox and the Blue Box Malware Campaign

b. Backdoor
    i. Syntax: OSINT Backdoor: <title from artifact>
    ii. Example: OSINT Backdoor: Purple Cheese Backdoor Strikes Windows
c. Botnet
    i. Syntax: OSINT Botnet: <title from artifact>
    ii. Example: OSINT Botnet: The Cool Kid Botnet Takes Over Gaming Machines
d. Ransomware
    i. Syntax: OSINT Ransomware: <title from artifact>
    ii. Example: OSINT Ransomware: Lock Your Stuff Ransomware Strikes Again
e. Tool
    i. Syntax: OSINT Tool: <title from artifact>
    ii. Example: OSINT Tool: Book Remote Access Trojan on the Rise in Asia

However, our example describes both a threat actor and a campaign. For titling purposes, we will use the first sub-category for titling schema standard's sake and record the representative descriptions using the appropriate galaxy tags found within the CMCF. Furthermore, like the creation of a Data Breach event, the Date is the date of event creation, Distribution is set to All Communities, Threat Level is set to Undefined, and Analysis is set to Completed.



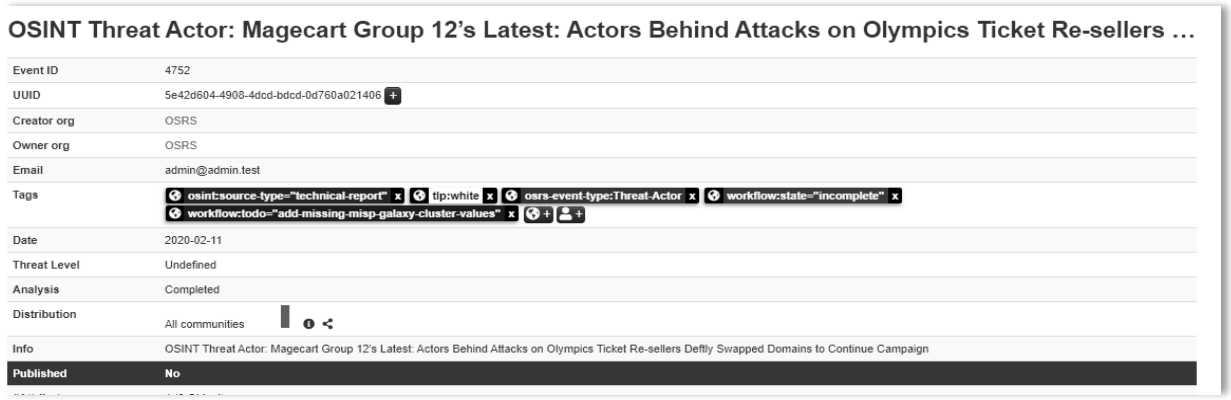
Once the information is submitted, we will be taken to the events main page. As with the data breach event, we will begin recording the event by adding the necessary event information via a template. Here, the template is much simpler than the data breach template. The only information that is initially added to the event is the Traffic Light Protocol Level (TLP), type of artifact, and the link to the source artifact. Sometimes there is a need to add additional links as the reporting artifact may not be the actual source of information but rather a summary based on the original source.

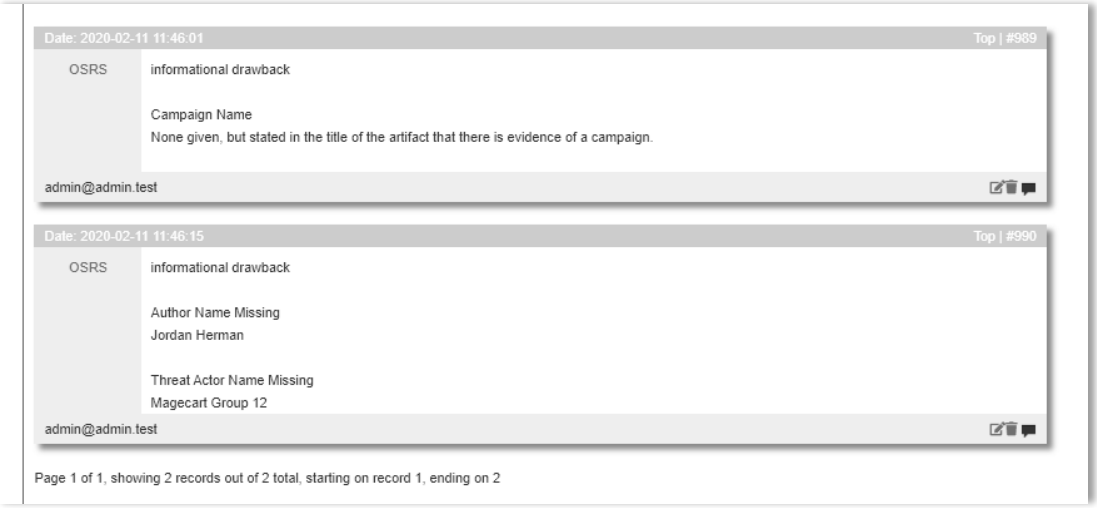
Now that the necessary information has been entered, we can begin deconstructing the information as well as annotation using a similar process to the one used during the recording of a data breach artifact. First, we will annotate the name of the threat actor, the name of the campaign, the reporting source, and the publishing author.

From our notes, we can see that the reporting source is *RiskIQ*, the publishing author is *Jordan Herman*, the threat actor is *Magecart Group 12*, but there is no name given to the campaign (though the term campaign is used in the title of the artifact). Therefore, we will not be attaching any campaign galaxy tag to the event. Instead, this now becomes an informational drawback. Based on the annotation, we can add the following tags from the CMCF section *General Artifact* to the event:

Artifact Type: Threat Actor
Artifact Type: Campaign
OSINT Source: RiskIQ Labs
Informational Drawback: MalTech Information Absent

Additionally, while adding the above galaxy tags to the event, we noticed that there were no tags within the CMCF collection to address the *Author Name*, as well as the name used for the threat actor *Magecart Group 12*. As previously performed during the Data Breach event, we will now add the two MISP workflow event tags reflecting *incomplete* and *missing galaxy cluster* references, as well as annotating the specific missing elements as notes attached to the event.





MalTech events divert from *General Artifact* tagging with regards to GDBI, GECI, U.S. State Index, ENISA DB SE Rating, and BLS industry coding attachments. Fundamentally, and by OSRS doctrinal philosophies, MalTech events will never be tagged with this type of information. One good example of doctrinal axioms for not tagging MalTech events with geolocation information is when sources report specific regions targeted by a particular MalTech entity. Myopically, this is from the viewpoint of the source who may or may not have all the necessary data to make this assertation. Are only these regions affected? Taking this factor into account adds to an element of inaccuracy in recording the data. Instead, this information is recorded in the textual commentary provided by the artifact as anecdotal to the entire artifact. Thus, this does not result in an informational drawback; instead, it is simply not considered relevant to the event.

Like recording a Data Breach artifact, MalTech artifacts present us the opportunity to walk through the general sections and sub-categories of the CMCF. Considering the following list, we have already recorded the *General Artifact* and *Author/Source* information. We can now proceed with the rest of the list; but, with one caveat: the format for which the information is presented.

1. General Artifact
2. Author/Source
3. Block-H
4. Combined Operational Framework
5. MITRE Pre ATK
6. MITRE ATK
7. SecTool
8. CWE

MalTech artifacts are elusive in their language. So much so that it typically helps the analyst to examine each section and annotate their findings based on the perspective of each CMCF category before actually breaking down the information and recording it on their respective tables. There is a propensity to discover a mixing and matching of relevant categorical

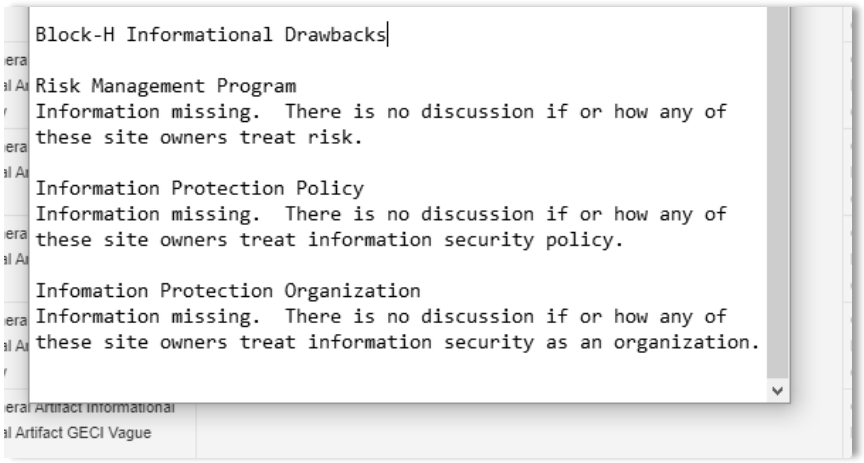
information within individual paragraphs. The opening paragraph of the artifact is an excellent example of how these various categories can be found within a single set of text.

> A recent blog post by Jacob Pimental and Max Kersten highlighted Magecart activity targeting ticket re-selling websites for the 2020 Olympics and EUFA Euro 2020, olympictickets2020.com and eurotickets2020.com respectively. These sites were compromised by a skimmer using the domain opendoorcdn.com for data exfiltration. With RiskIQ data, our researchers built on the previous reporting to identify more skimming domains used by the attackers, as well as additional compromised sites. RiskIQ can also now attribute all these compromises to Magecart Group 12.

There are two findings presented in the text. The first is that the threat actor targeted publicly facing websites. The second is that the actor used automated domain data exfiltration. Just taking these two items into account, we can assume that we can add the following tags at some point during the process. However, we will not do that at this time. Instead, these are merely examples of how multiple categories can be presented in a single paragraph.

- Block-H
    - Acquisition and Development: 10.b Securing Application Services on Public Networks
        - The organization failed to secure its publicly facing data processing application against this type of threat actor and the attack
- Combined Operational Framework
    - Human Threat: ENISA Inadequate Design and Planning or Improper Adaptation
        - The organization failed to secure its publicly facing data processing application against this type of threat actor and the attack
- MITRE Pre ATK
    - Technical Weakness Identification: T1293 Analyze Application Security Posture
        - The adversary did sufficient research into the application posture to know where the application was weakest and allowed for the highest success rate in an attack.
- MITRE ATK
    - Initial Access: Linux T1190 Exploit Publicly Facing Application
    - Initial Access: Windows T1190 Exploit Publicly Facing Application
        - The application may or may not reside on a Linux or Windows host. So, we will use both here assumptively. Regardless, the application was publicly facing and was compromised to conduct the rest of the attack.

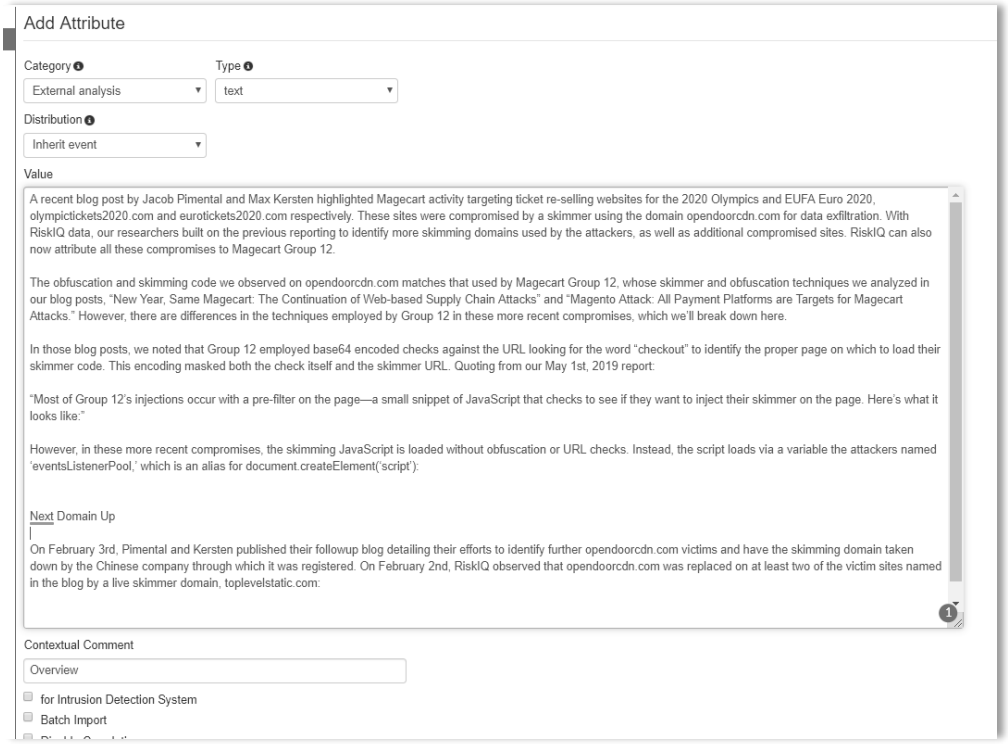
Along with these pieces of necessary information, several informational drawbacks were quickly identified. For example, many data points related to Block-H, but much of the data did not convey how any of the site owners treated cybersecurity from a risk perspective. Some of the quick annotations regarding this are as follows:



Honestly, a lot of deconstruction of an artifact relies on the familiarity an analyst has with the various ways to describe data using frameworks like MITRE or CWE. Conjoined with familiarity is an analyst's ability to apply critical thinking to the data and derive assumptions based on what is presented by the artifact. Familiarity and critical thinking are skills that mature over time and are refined by going through the process of artifact deconstruction repeatedly.

Consequently, this reliance on analyst experience creates an inherent variance of accuracy in artifact recording; meaning, a seasoned analyst, should be able to derive more and accurate information from an artifact than a junior analyst who is new to the process. Nonetheless, we have taken the time to complete the annotation of the artifact. We can now begin to go through the tables for tag determination using the process that was previously applied to the tagging of a Data Breach artifact.

However, before we begin adding tags, and after we have fully annotated the artifact, we must add the body of text from the artifact to the event. To do this, we simply navigate to the left-hand side of the main event page to add an attribute to the event. In this case, we would similarly add the text like that shown in the image below. Simply copy and paste the text from the artifact to the into the *Value* portion of the attribute page. Once this is step is completed, we can proceed with tagging. Also, notice how this attribute is set up. *Category* is set to *External analysis*, the *Type* is set to *text,* and the *Contextual Comment* is annotated with the term *Overview.*



Section 2: Adding Block-H Tags to the Event

Similar to the Data Breach artifact, we are looking at where, if at all, within the Block-H portion of the CMCF, this artifact presents data that describes instances whereby Block-H either was not represented, failed, or in the case of MalTech, could fail.

Block-H Sub-Category	Were any of the sub-categories violated, or did any of the sub-categories relate to a breakdown of the CIA triangle?	Are there instances or opportunities for informational drawbacks?
Risk Management Program	Risk management program was never addressed in the artifact.	Information absent.
Information Protection Policy	Information Protection Policy was never addressed in the artifact.	Information absent.
Information Protection Organization	Information Protection Organization was never addressed in the artifact.	Information absent.
Human Resources	Clearly, the sites were vulnerable to this type of threat actors attack methodology	Information absent.
Asset Management	The artifact does not address if any asset management was applied at all by the website owners.	Information absent.
Access Control	The vulnerability within the websites allowed for sensitive information to be accessed by the adversary. The attack was also directed at the websites source code as a code injection type of attack.	
Cryptography	There was no mention within the artifact regarding the use of cryptographic controls.	Information absent.
Physical/Environmental Protection	The attack was performed on publicly facing websites and was not the result of any physical or environmental threat.	Information not relevant.

Operations	Any change management review did not catch the code injection. The injected JavaScript is a form of malware that was not protected against by the website owners. The websites did not capture logs detailing this anomalous behavior. JavaScript was injected into operational systems. The vulnerability was left unmanaged by the website owners. The JavaScript was not restricted from loading on operational systems.	
Communications	The websites did not secure the information from anomalous or deviant data transfer.	
System Acquisition/Development	It was not clear as to what acquisitional security requirements were applied to the websites when developed/purchased. However, the application was not secured on publicly available services. The vulnerability left the transactions in a vulnerable, unprotected state. The installation of the JavaScript was not restricted on the website. The websites, due to the exposure of their vulnerability, were not securely engineered. The sites were accepted in a vulnerable state before deployment to publicly facing services.	Information absent.
Supply Chain	There is no mention within the artifact if 3rd Party services maintain any of the website infrastructures.	Information absent.
Incident Management	It is unclear after the artifact was published to what extent, if any, the website owners addressed the information relating to their vulnerable websites.	Information absent.
Business Continuity	The artifact does not describe data in a way that relates to this category.	Information not relevant.
Compliance	The organization failed, through the use of a vulnerable website, was unable to protect sensitive information contained within transactional records. Additionally, the vulnerability that allowed the attack to occur was either accepted during a technical compliance review, no compliance review was conducted, or the information was never detected during a technical inspection of the website's security.	
Privacy	The artifact does not discuss if a privacy policy was presented on any of the affected websites, nor if any of the owning organizations had sufficient or existent privacy policies posted.	Information absent.

From the table we can attach the following Block-H galaxy tags to the event:

Human Resources: 03.c Management Responsibilities
Human Resources: 03.d Information Security Awareness, Education, and Training
Access Control: 05.j Information Access Restriction
Access Control: 05.n Access Control to Program Source Code
Operations: 08.b Change Management
Operations: 08.e Controls Against Malware
Operations: 08.g Event Logging
Operations: 08.k Installation of Software on Operational Systems
Operations: 08.l Management of Technical Vulnerabilities
Operations: 08.m Restrictions on Software Installation
Communications: 09.d Information Transfer
Acquisition and Development: 10.b Securing Application Services on Public Networks
Acquisition and Development: 10.c Protecting Application Services Transactions
Acquisition and Development: 10.h Restrictions on Changes to Software Packages
Acquisition and Development: 10.i Secure System Engineering Principles
Acquisition and Development: 10.m System Acceptance
Compliance: 14.c Protection of Records
Compliance: 14.h Technical Compliance Review

Now that we have attached all the relevant Block-H galaxy tags, we can assign any notes regarding the discovered informational drawbacks. The image below shows the annotation of informational drawbacks attached to the event.

Informational Drawback tags regarding Block-H attached to the event are as follows:

Informational Drawback: Block-H Information Absent
Informational Drawback: Block-H No Relevant Category

Section 3: Adding Combined Operational Framework Tags to the Event

From here, we move on to attaching relevant Combined Operational Framework tags. In the same manner, as the Data Breach artifact deconstruction, we can use a table to help organize our thoughts into a workable format that allows us to justify the tags we will attach to the event.

Combined Operational Framework Sub-Category	Were any of the sub-categories violated, or did any of the sub-categories relate to a breakdown of the CIA triangle?	Are there instances or opportunities for informational drawbacks?
Environmental Threat	The threat described within the artifact does not relate to an environmental threat	
Failure Threat	The threat described within the artifact does not relate to a threat of system failure	
Organizational Threat	The threat described within the artifact does not specify an organizational threat	
Existential Threat	The artifact does not describe the data as an existential threat	
Human Threat	The websites were improperly designed to prevent this type of attack. The data was	

	stolen electronically. The activity caused by the vulnerability was not detected in system logs. The website was not configured to prevent this vulnerability from being exploited.	
Nefarious Threat	The information was *skimmed* or intercepted by the adversary. The vulnerability allowed for the injection of malicious code. The JavaScript was an unauthorized piece of software. The JavaScript compromised legitimate websites to fool victims into a false sense of legitimacy. By injecting JavaScript, the website's source code was altered.	

From the results of the table, we can attach the following tags from the CMCF *Combined Operational Frameworks* category:

Informational Drawback: Block-H Information Absent
Human Threat: ENISA Inadequate Design and Planning or Improper Adaptation
Human Threat: HITRUST LIM4 Theft
Human Threat: HITRUST LUM11 Lack-of or Insufficient Logging
Human Threat: HITRUST LUM15 System Configuration Errors
Human Threat: HITRUST LUM7 Improperly Designing Information Systems
Nefarious Threat: ENISA Interception of Information
Nefarious Threat: ENISA Malicious Code, Software, or Activity
Nefarious Threat: ENISA Unauthorized Installation of Software
Nefarious Threat: HITURST LIN16 HTML Script Injection
Nefarious Threat: HITRUST LIN21 Malicious Code Execution
Nefarious Threat: HITRUST LIN25 Masquerade or Pretexting
Nefarious Threat: HITRUST LIN3 Alteration of Software
Nefarious Threat: HITRUST LIN7 Code Injections

Section 4: Adding MITRE Pre ATK Tags to the Event

We are now at a point where we can attach tags to the event that relate to the MITRE Pre ATK framework. We will continue to use the table to keep ourselves organized as we move through this procedural process.

MITRE Pre ATK Framework Sub-Category	Were any of the sub-categories of the MITRE Pre ATK Framework sufficiently described within the artifact.	What type of information drawbacks may exist for this sub-category?
Priority Definition Planning	The actors did define their plan and created a set of priorities before launching this type of attack. There were distinct leadership areas of interest. It can be assumed that the gaps were identified. It is unclear how a cost-benefit analysis was conducted, so this is a bit vague. Due to the vulnerability targeted, it can be assumed that this was a part of the adversary's KIT/KIQ. Due to the vulnerability targeted, it can assumptively there was some sort of overall plan, but how strategically developed this plan was is	Information vague (CBA) Information vague (Strategic Planning)

	vague. To deliver the JavaScript, there must have been an implementation plan.	
Priority Definition Direction	Due to the nature of the attack there it can be assumed that KITs were submitted in some fashion, KITs were assigned in some manner, requirements for KITs were received in some way, and requirements for tasks were created in some style.	
Target Selection	From the information reported within the artifact, it is clear that strategic targets were at some point determined, operational elements were determined in some fashion, highest level tactical elements were identified in some manner, as the approach or attack vector. However, it is unclear if there were tactical elements below the primary tactical component, even though there was more than one website affected in the attack.	Information vague (unknown levels of tactical elements)
Technical Information Gathering	The sites were publicly facing so easy to develop OSINT information on them. The sites themselves were attacked, so there was no real need to conduct social engineering per se or dig through job openings to find the websites. IP addresses and domains were identified. Due to the nature of the attack obtaining domain owner information was probably not required. The attack was public, so no need for network mapping unless to pivot to through the associated links from the top-level website URL. Website defenses were identified.	Information vague (passive scanning)
People Information Gathering	The attack was targeted at websites and did not include the need to target people.	
Organizational Information Gathering	The attack was targeted at websites and did not include the need to target the owning organization directly.	
Technical Weakness Identification	The adversary was able to identify websites that contained weaknesses or vulnerabilities that allowed for the JavaScript to be injected into the websites source code and execute.	
People Weakness Identification	The attack was targeted at websites and did not include the need to target people.	
Organizational Weakness Gathering	The attack was targeted at websites and did not include the need to target the owning organization directly.	
Adversary OpSec	A proxy service was used as a part of the redirect of intercepted data. Redirect domains were not entirely private or anonymized. Adversary infrastructure was hosted on 3rd Party services. Pieces of the JavaScript, which is a part of the adversary's software infrastructure, were obfuscated. Standard and high protocols were used.	
Establish/Maintain Infrastructure	Redirect domains were purchased. 3rd Party infrastructure services were used. The attack required some level of technical resources, and these were obtained to conduct the attack. 3rd Party infrastructure required additional configuration to enable the attack. Redirect sites used Let's Encrypt certificates. Though there was more than one infrastructure identified in the artifact, it is unclear if any were considered primary or backup resources.	Information Vague (backup infrastructure)
Persona Development	The attack was targeted at websites and did not include the need to develop a persona.	
Build Capabilities	It is unclear if the JavaScript was created or purchased by the adversary, only that a	Information Vague (create or obtain payloads).

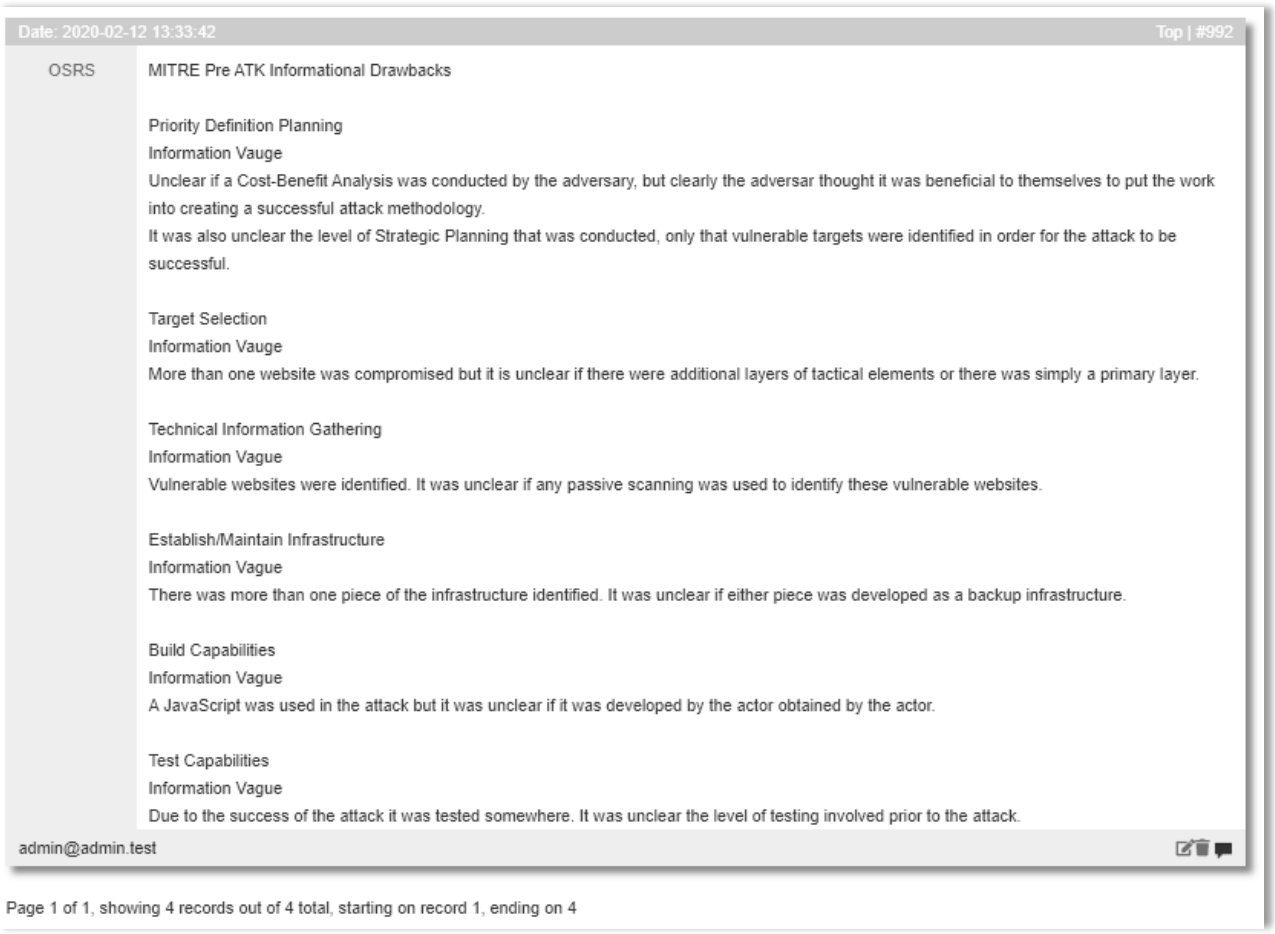
	JavaScript that exploits this type of website vulnerability explicitly was used. Delivery systems were built as well as identifying requirements to build these capabilities.	
Test Capabilities	The JavaScript clearly works, so it must have been tested somewhere. Still, that information is vague or absent from the artifact, and only enough information is provided to produce a general assumption.	Information Vague (testing capabilities)
Stage Capabilities	Capabilities required staging before the attack.	

From the results of the table, we have enough information to tag the event with the following MITRE Pre ATK galaxy tags:

Priority Definition Planning: T1224 Assess Leadership Areas of Interest
Priority Definition Planning: T1225 Identify Gap Areas
Priority Definition Planning: T1227 Develop KITs or KIQs
Priority Definition Planning: T1228 Assign KITs or KIQs Into Categories
Priority Definition Planning: T1229 Assess KITs or KIQs Benefits
Priority Definition Planning: T1232 Create Implementation Plan
Priority Definition Direction: T1237 Submit KITs, KIQs, and Intelligence Requirements
Priority Definition Direction: T1238 Assign KITs, KIQs, and (or) Intelligence Requirements
Priority Definition Direction: T1239 Receive KITs, KIQs, and Determine Requirements
Priority Definition Direction: T1240 Task Requirements
Target Selection: T1241 Determine Strategic Target
Target Selection: T1242 Determine Operational Environment
Target Selection: T1243 Determine Highest Level Tactical Element
Target Selection: T1245 Determine Approach or Attack Vector
Technical Information Gathering: T1247 Acquire OSINT Data Sets and Information
Technical Information Gathering: T1250 Determine Domain and IP Address Space
Technical Information Gathering: T1256 Identify Web Defensive Services
Technical Weakness Gathering: T1287 Analyze Data Collected
Technical Weakness Gathering: T1288 Analyze Architecture and Configuration Posture
Technical Weakness Gathering: T1291 Research Relevant Vulnerabilities and (or) CVEs
Adversary OpSec: T1304 Proxy and (or) Protocol Relays
Adversary OpSec: T1307 Acquire and (or) Use 3rd Party Infrastructure Services
Adversary OpSec: T1319 Obfuscate or Encrypt Code
Adversary OpSec: T1321 Common, High Volume Protocols and Services
Establish/Maintain Infrastructure: T1328 Buy Domain Name
Establish/Maintain Infrastructure: T1329 Acquire and (or) Use 3rd Party Infrastructure Services
Establish/Maintain Infrastructure: T1335 Procure Required Equipment and Software
Establish/Maintain Infrastructure: T1339 Install and Configure Hardware, Network, and Systems
Establish/Maintain Infrastructure: T1337 SSL Certificate Acquisition for Domain
Build Capabilities: T1347 Build and Configure Delivery Systems
Build Capabilities: T1348 Identify Resources Required to Build Capabilities
Stage Capabilities: T1362 Upload, Install, and Configure Software or Tools

We can also see from examining the artifact that there were a few informational drawbacks identified as well. Using the results from the table, we can add the following informational drawback galaxy tags related to the MITRE Pre ATK framework to the event; as well as add annotation of these drawbacks to the event:

Informational Drawback: MITRE Pre ATK Artifact Vague (Build Capabilities)
Informational Drawback: MITRE Pre ATK Artifact Vague (Establish & Maintain Infrastructure)
Informational Drawback: MITRE Pre ATK Artifact Vague (Priority Definition Planning)
Informational Drawback: MITRE Pre ATK Artifact Vague (Target Selection)
Informational Drawback: MITRE Pre ATK Artifact Vague (Technical Information Gathering)



Section 5: Adding MITRE ATK Tags to the Event

Now that we have the MITRE Pre ATK galaxy tags attached to the event and the informational drawbacks annotated, we can move on to the next CMCF category, the MITRE ATK framework. We will use the same methodology for recording notes into a table to stay organized through this procedure.

MITRE ATK Framework Sub-Category	Were any of the sub-categories of the MITRE ATK Framework sufficiently described within the artifact.	What type of information drawbacks may exist for this sub-category?
Initial Access	The adversary exploited publicly available websites. However, it is unclear the type of operating system that was used to host the websites.	Information absent (no reference to the operating system that hosted the website)
Execution	Looking back over the artifact, we see that the script uses the following command, which does not require a user to launch the JavaScript. *script async=*	
Persistence	The JavaScript will remain on the page unless removed by the website's owner. Additional persistence information that might align to the MITRE ATK framework is missing.	Information absent

Privilege Escalation	Adding the JavaScript to the websites source code must have required some sort of privilege. However, this information is missing from the artifact.	Information absent
Defense Evasion	Some versions of the JavaScript used obfuscation to evade discovery. Some versions of the JavaScript checked for specific flags before executing the JavaScript.	
Credential Access	Adding the JavaScript to the website's source code must have required some sort of access requirement. However, this information is absent from the artifact.	Information absent
Discovery	Knowing which information to intercept must have been discovered, but this information is missing from the artifact.	Information absent
Lateral Movement	No lateral movement was observed	
Collection	Data was automatically collected via the injected JavaScript.	
Command and Control	The JavaScript was loaded from other .com sites, but it does not clearly state what protocols were used to deliver the JavaScript (i.e., standard/non-standard protocols)	Information vague
Exfiltration	The JavaScript was used to automate the exfiltration of the data. However, the data is also ambiguous as to if the data was encrypted or if any other exfiltration method was used.	Information vague
Impact	There is no relevant category for this type of attack.	No relevant category

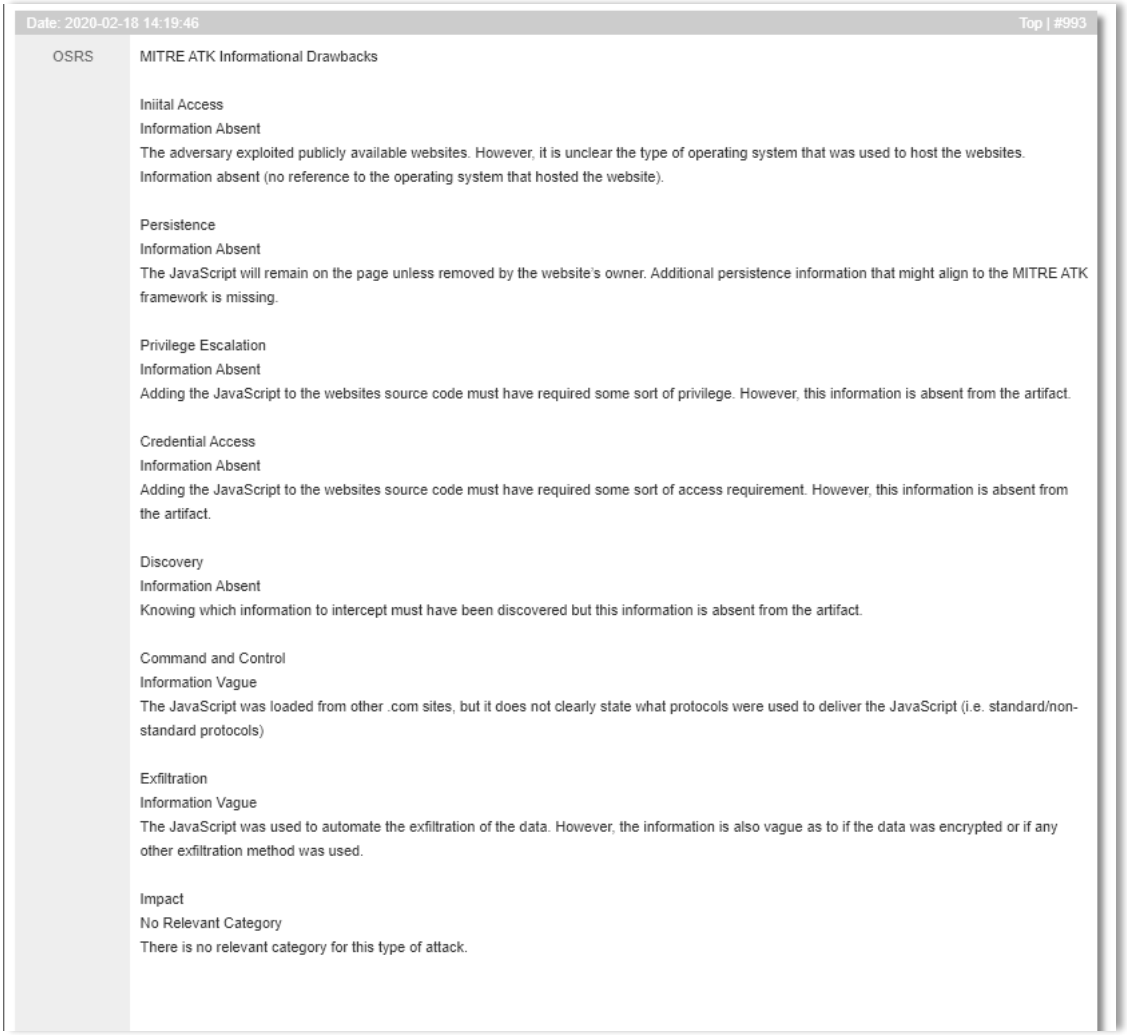
From the results in the table we can confidently attach the following MITRE ATK galaxy tags to the event:

Initial Access: Linux T1190 Exploit Public Facing Application
Initial Access: Windows T1190 Exploit Public Facing Application
Execution: Linux T1064 Scripting
Execution: Windows T1064 Scripting
Defense Evasion: Linux T1027 Obfuscated Files or Information
Defense Evasion: Linux T1480 Execution Guardrails
Defense Evasion: Windows T1027 Obfuscated Files or Information
Defense Evasion: Windows T1480 Execution Guardrails
Exfiltration: Linux T1020 Automated Exfiltration
Exfiltration: Windows T1020 Automated Exfiltration

As we can see from the table, this artifact leaves out or is vague in its explanation for several MITRE ATK framework sub-categories. Nonetheless, we must now include the following informational drawbacks to the event:

Informational Drawback: MITRE ATK Information Absent (Initial Access)
Informational Drawback: MITRE ATK Information Absent (Persistence)
Informational Drawback: MITRE ATK Information Absent (Privilege Escalation)
Informational Drawback: MITRE ATK Information Absent (Credential Access)
Informational Drawback: MITRE ATK Information Absent (Discovery)
Informational Drawback: MITRE ATK Information Absent (Credential Access)
Informational Drawback: MITRE ATK Information Vague (Command and Control)
Informational Drawback: MITRE ATK Information Vague (Exfiltration)
Informational Drawback: MITRE ATK No Relevant Category (Impact)

And lastly, we would add the specific notes regarding each informational drawback relating to the MITRE ATK framework to the event as well.


Date: 2020-02-18 14:19:46 Top | #993

OSRS

MITRE ATK Informational Drawbacks

Iniital Access
Information Absent
The adversary exploited publicly available websites. However, it is unclear the type of operating system that was used to host the websites. Information absent (no reference to the operating system that hosted the website).

Persistence
Information Absent
The JavaScript will remain on the page unless removed by the website's owner. Additional persistence information that might align to the MITRE ATK framework is missing.

Privilege Escalation
Information Absent
Adding the JavaScript to the websites source code must have required some sort of privilege. However, this information is absent from the artifact.

Credential Access
Information Absent
Adding the JavaScript to the websites source code must have required some sort of access requirement. However, this information is absent from the artifact.

Discovery
Information Absent
Knowing which information to intercept must have been discovered but this information is absent from the artifact.

Command and Control
Information Vague
The JavaScript was loaded from other .com sites, but it does not clearly state what protocols were used to deliver the JavaScript (i.e. standard/non-standard protocols)

Exfiltration
Information Vague
The JavaScript was used to automate the exfiltration of the data. However, the information is also vague as to if the data was encrypted or if any other exfiltration method was used.

Impact
No Relevant Category
There is no relevant category for this type of attack.

MalTech

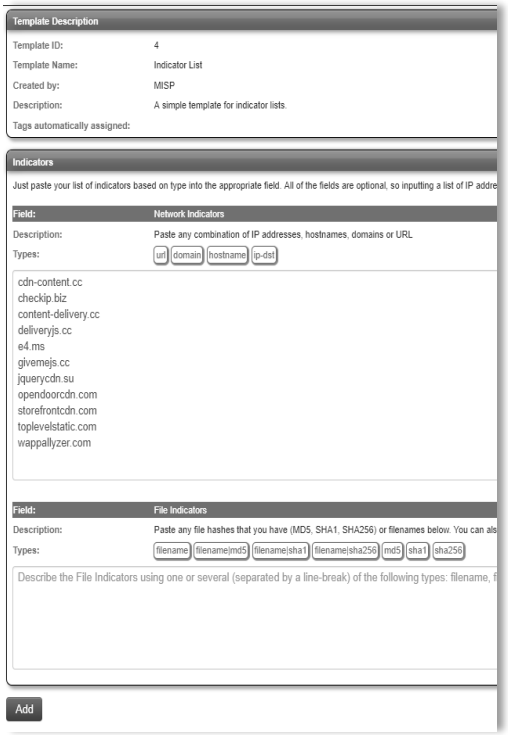
Continuing to move through the process for recording a MalTech OSINT artifact, we will now attach any of the relevant CMCF or MISP tags that would be appropriate from the event. However, if we look back at the beginning of this process, we would find that there are no relevant tags within the CMCF or MISP database regarding the threat actor. Therefore, this step in the process would be annotated with the missing information, and the built-in MISP taxonomy for workflow would be attached to the event to represent this missing information. Since we have already completed this in a different section of the process, there is no need to reiterate the instructions on how to handle this type of missing information.

Section 6: Adding SecTool Tags to the Event

Sometimes a MalTech artifact will call out the use of a SecTool. However, this artifact did not. Therefore, there is no need to attempt to annotate the event or tag the event with informational drawbacks regarding missing (or vague) SecTool information.

Section 7: Adding CWE Tags to the Event

This portion of the process will most certainly be left to analyst conjecture. The websites were vulnerable to script injection. Using the search function inherent in the MISP map capability set, we find that there are CWE's that directly relate to code injection. Furthermore, since we know that the injected code is JavaScript, a simple search Google search for [*JavaScript and CWE*] results in the following CWE: *CWE-83: Improper Neutralization of Script in Attributes in a Web Page*. At this point, either tag or both tags would be relevant to the event. However, the CMCF does not contain a galaxy tag for CWE-83. For this specific example, we will add *CWE-94 Improper Control of Generation of Code (Code Injection)* to the event.

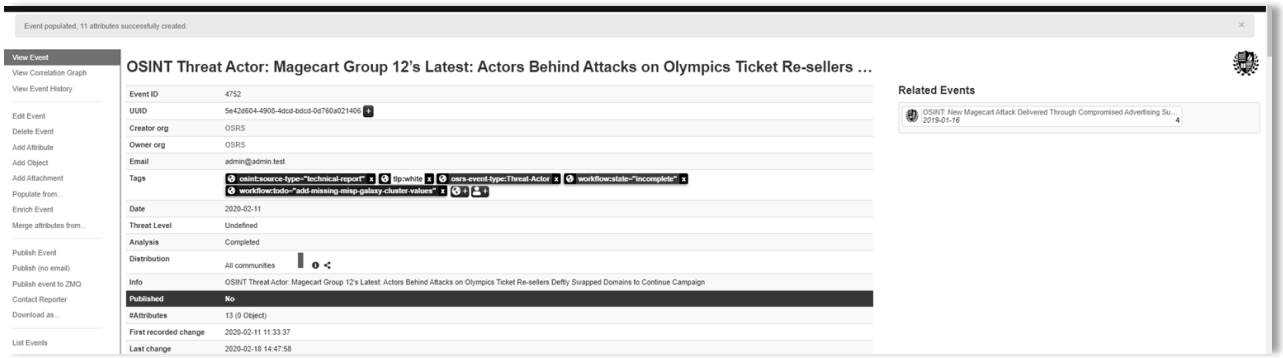


Section 8: Adding Indicators of Compromise (IoC) to the Event

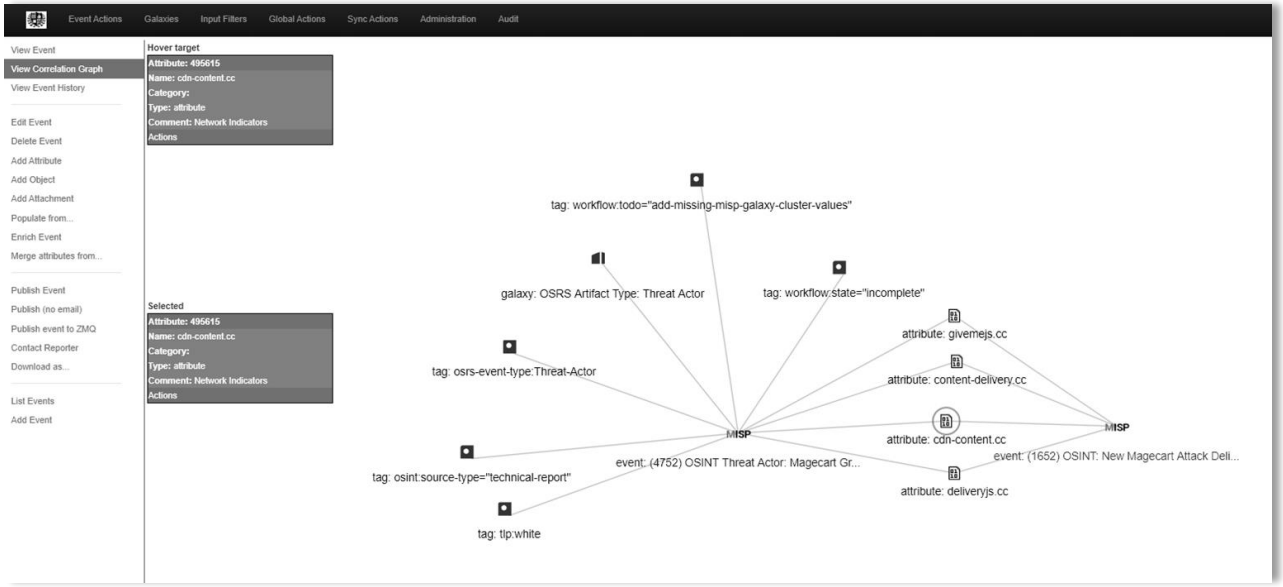
Some MalTech OSINT artifacts will contain information relating to Indicators of Compromise, or IoCs. There is no exact amount of IoC information per artifact. Some MalTech artifacts will only have a small number of IoCs, while others will contain hundreds of IoCs. Typically, IoCs take the form of IP address, domain, URL, file name, and file hash information relating to the information contained within the artifact. For example, the addition of IoCs to this specific artifact would be IoCs that relate to *Magecart Group 12*. There are a few ways we can add IoCs to the event. In this example, we will add them via a bulk method using a built-in template from MISP. Since many of the indicators presented in the

artifact are simply domains, we will use the MISP template titled *Indicator List.* Again, to do this, we simply need to paste the IoCs in the appropriate text box within the template. Once completed, submit and review the indicators before attaching them to the event.

Sometimes, and immediately after attaching IoCs to an event, MISP will make a correlation or relationship to another event stored within MISP. This correlation means that there are attributes in other events that are duplicates. This corollary function is to help analysts understand the structure of information and how the information contained within an event stands alone, but also as a community of data that has natural relationships share between other events. Any event related to the current event we are working is found as a list of events in the upper right-hand corner of the main event page as seen in the image below.



Another way to look at the relationships between events is with MISP's built-in correlation graph capability. One way to view this type of graph within MISP is to simply click the link located on the left-hand side of the main event page title appropriately *View Correlation Graph.* The following image is what an analyst would see if they had clicked the link to the graph.



Section 9: Adding Artifact Attachments to the Event

Last but not least, some MalTech OSINT artifacts contain images that might be useful to other analysts when attempting to understand or synthesize the material within an event. To add pictures to an event, we must add them as an attachment. To add an attachment, we simply need to navigate to the attachment page via the *Add Attachment* link located on the left-hand side of the main event page. There is no clear guidance as to how much imagery should be attached to the event. Each analyst will have a level of bias when determining how much and which images to add as an attachment.

Furthermore, some types of attachments might be in the form of PDF or other text files. There are times when the initial or source artifact that an analyst finds contains a link to a more in-depth report. It is up to each analyst on how to handle this type of situation. From the perspective of time management, some of these additional resources contain more information than worth recording in an event. Therefore, it is simply easier to add these more in-depth resources to the event as an attachment.

Chapter Summary

In this chapter, we learned the basics of recording a MalTech type of event. We have also seen how analyst bias enters into the recording process as the artifacts themselves contain informational drawbacks that do not layout where, in any framework, the information would be relevant. Analyst experience plays a large part in how their bias is applied to the event. In a later chapter, we look at data mining the information in MISP using Splunk and MISP42. We will begin to see how this bias and lack of precise details portrayed in an artifact affect the results, and ultimately the statistics that can be performed on the data for decision support.

Chapter Acronyms

ATK: MITRE ATT&CK Framework

BLS: U.S. Bureau of Labor Statistics

CWE: Common Weakness Enumeration

ENISA: The European Union Agency for Cybersecurity

GDBI: General Data Breach Index

GECI: Government Entity Classification Index

IOC: Indicator of Compromise

MISP: Malware Information Sharing Platform

OPSEC: Operational Security

OSINT: Open Source Intelligence

TLP: Traffic Light Protocol

Chapter References

Common Weakness Enumeration. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Weakness_Enumeration&oldid=918897150

CWE - CWE-94: Improper Control of Generation of Code ('Code Injection') (3.4.1). (n.d.). Retrieved February 18, 2020, from https://cwe.mitre.org/data/definitions/94.html

ENISA. (n.d.). Retrieved February 4, 2020, from https://www.enisa.europa.eu/

Home | U.S. Bureau of Labor Statistics. (n.d.). Retrieved February 4, 2020, from https://www.bls.gov/

Magecart Group 12's Latest: Deftly Swapping Domains to Continue Attacks. (n.d.). Retrieved February 18, 2020, from https://www.riskiq.com/blog/labs/magecart-group-12-olympics/

MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing (formely known as Malware Information Sharing Platform). (n.d.). Retrieved February 4, 2020, from https://www.misp-project.org/

MITRE ATT&CK™. (n.d.). Retrieved February 18, 2020, from https://attack.mitre.org/

Open-source intelligence. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open-source_intelligence&oldid=937269341

Operations security. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Operations_security&oldid=922796009

Traffic Light Protocol (TLP) Definitions and Usage | CISA. (n.d.). Retrieved February 18, 2020, from https://www.us-cert.gov/tlp

Chapter 4: Recording Fault and SecTool OSINT Artifacts

Chapter 4 Introduction

In this chapter, we look at the recording of Fault and SecTool category OSINT artifacts. As seen in Chapter 1, the parent category of *Fault* has two child categories: *Vulnerabilities* and *Exploits.* Though mutually exclusive, there are times, like with the *MalTech* category, where information from an artifact can comingle between the two child categories. For example, it is not uncommon for an exploit artifact to have been published that also contains associated CVE information.

Additionally, as we have seen with both the Data Breach and MalTech categories, artifacts can tend to have a large number of tags attached to the event. Purposeful in intention but can become cumbersome when applied to the *Fault* category. Often the information contained within a *Fault* artifact is merely absent, or at best, vague concerning the CMCF mapping. MITRE is one example of where there simply is no information provided by the artifact that aligns with the framework. At this point, it is up to the individual analyst or organization to determine how cumbersome or agile they wish to perform with the documentation of information drawbacks found in *Fault* artifacts.

The goal of this chapter is to present a set of example information that gives an analyst an essential foundation for recording these types of artifacts. This chapter will not go through extended examples of how to annotate or organize an analyst's workflow. Instead, we will be using Chapters 1 and 2 as building blocks. The process for recording vulnerability or exploit information is not that different than recording any other artifact.

Recording SecTool and Fault Artifacts

Section 1: Artifact Selection and Event Creation

This chapter will use three reference artifacts collected online. For the exploit reference, we will be using *Nanometrics Centaur 4.3.23 - Unauthenticated Remote Memory Leak* found on the online database Exploit DB. The vulnerability example, *CVE-2020-0728: Windows Modules Installer Service Information Disclosure Vulnerability*, has also been obtained online from the open resource *Bugtraq.* Lastly, we will use *Suricata IDPE 5.0.2* collected from the website *Packet Storm.* Each of these source sites can be automatically searched and placed within a queue by using any RSS or web scraping tool.

Beginning the events starts by using the same process from the previous chapters. All general event creation settings will remain the same. What is different will be the title given to each event. The following titles will be used for each separate artifact:

OSINT SecTool: Suricata IDPE 5.0.2

OSINT Vulnerability: CVE-2020-0728: Windows Modules Installer Service Information Disclosure Vulnerability

OSINT Exploit-DB: Nanometrics Centaur 4.3.23 - Unauthenticated Remote Memory Leak

Once the event has been created, we can add in the initial information from each artifact into their respective events. This can be done using templates or by individual attributes. Each event will have a minimum set of taxonomy tags that declare the event's TLP and a link that references the original source of the OSINT artifact. In the case of exploits and vulnerabilities, if there is a CVE presented in the artifact, this would be recorded as well. The built-in MISP template *Malware Report* contains a section for recording CVE information.

Like recording Data Breach and MalTech information, it would be this portion of the process where an analyst would enter in the text for each event. In the case of SecTool artifacts, we would record any relevant textual information regarding what the tool does or what updates have been applied to the tool. Using this specific SecTool example, we would record the updates as posted on the *Packet Storm* website. For vulnerability artifacts, we would record all textual information regarding the vulnerability description, how the vulnerability is triggered, and possible published fixes for the vulnerability. Lastly, for exploit artifacts, we would record the actual exploit code, if presented in the artifact, as a textual attribute to the event.

Section 2: Adding General Artifact and Author/Source Tags to the Event

In most cases, if not all cases, there will not be many tags from the CMCF General Artifact category that aligns with these types of artifacts. This limitation is deficient by design, as the General Artifact category only describes the artifact's information in distinct general terms. Examining the CMCF General Artifact category, we can see that only the sub-category of *Artifact Type* applies to any of these events. Along with the artifact type, we can also add in the author sources tags for each of the artifacts. As in previous chapters, we have seen that there are times when a relevant tag is not present within the CMCF. Therefore, we must apply event tags that denote what is missing as well as annotate this information within the individual event's *Notes* section as done before.

We will use the following tags for the relevant corresponding artifact.

SecTool

OSRS Artifact Type: Security Tool
OSRS SecTool: Suricata
OSRS SecTool Type: Network Intrusion Detection and Prevention Engine (IDPE)

**Packet Storm is missing from the CMCF, and so we will mark the event as incomplete. Furthermore, the author of the artifact is absent. This absence is typical because the author is an organization rather than a pen name or actual author name. For this, we will not annotate any informational drawback.*

Vulnerability

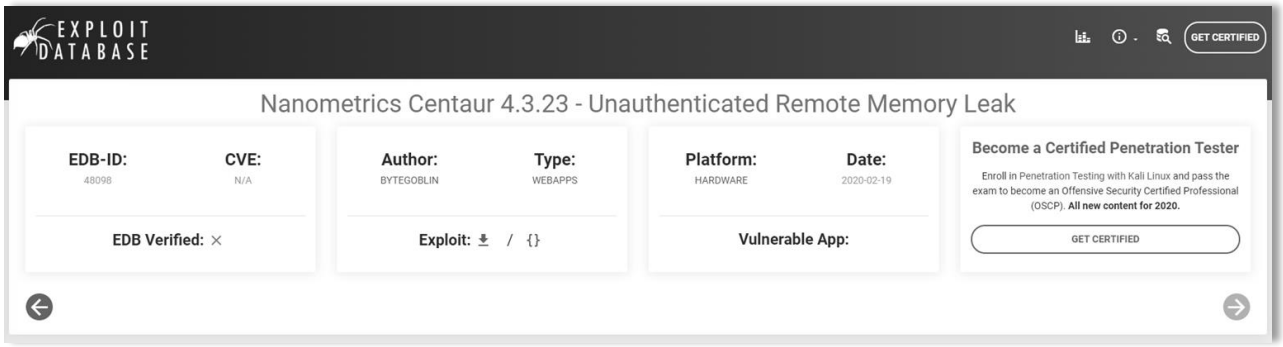
OSRS Artifact Type: Vulnerability
OSRS Vulnerability Source: Bugtraq
OSRS CVE Author Index: Imre Rad

Exploit

OSRS Artifact Type: Exploit
OSRS Vulnerability Source: exploit-db (The Exploit Database)
OSRS Exploit DB: Type WebApps
OSRS Exploit DB Verification: Not Verified
OSRS Exploit DB Platform: Hardware

**The author for this artifact is not contained within the CMCF, and therefore we must mark the event incomplete and annotate the event with the missing information.*

This point in the chapter is an excellent spot to stop and discuss the examination of these artifact types. Previously it was mentioned that if there is CVE information contained within an artifact that this information is recorded as well. The CVE information in the vulnerability artifact is easy to find. It's clearly stated in the title of the artifact ***CVE-2020-0728****: Windows Modules Installer Service Information Disclosure Vulnerability*. However, when we look at the web page that displays the exploit artifact information, we see that the website has omitted the CVE information.



Yet, as we begin to look over this particular artifact for annotation purposes, we see that the CVE, as well as the CWE information, is available.

```
#
# Tested on:
#   Jetty 9.4.z-SNAPSHOT
#
# Vulnerability discovered by:
#   byteGoblin @ zeroscience.mk
#
#
# Advisory ID: ZSL-2020-5562
# Advisory URL: https://www.zeroscience.mk/en/vulnerabilities/ZSL-2020-5562.php
#
# Related CVE: CVE-2015-2080
# Related CWE: CWE-532, CWE-538
#
# 10.02.2020
#
```

Section 3: Adding Block-H, Combined Operational Framework, and MITRE Tags to the Event

These frameworks were not intentionally developed to describe artifacts that contain *SecTool* or *Fault* type information. Due to the purposeful construction of these frameworks, attempts at attaching tags that align to these frameworks tend to create bloat with informational drawback tagging. It is not to say that we cannot attach tags from these CMCF categories to any of these examples; it is that the information at that point becomes a matter of conjecture and inference. More so than when attaching tags from these clusters to events derived from artifacts from the other parent CMCF categories.

Examination of the SecTool artifact reveals that Suricata is a Network Intrusion Detection and Prevention Engine (IDPE). These types of security tools are typically placed within an organization's network in such a way as to intercept and inspect network traffic for anomalous; or malicious traffic behavior patterns. By instituting an IDPE, an organization by inference is, in some way, addressing security as an organization as well as an operation. However, an IDPE is not explicitly called out in any of these frameworks. From the viewpoint of Block-H, an IDPE

would help to secure network services, and so, therefore, the Block-H tag *HITRUST CSF Communications: 09.b Security of Network Services* could be attached to the event.

Examination of both the vulnerability and exploit artifacts result in similar reliance on analyst bias and organizational goals for exact tagging of a Fault type event. Again, using the CMCF Block-H category as a referential example, attaching the tag *HITRUST CSF Operations: 08.l Management of Technical Vulnerabilities* is not an inappropriate tag to connect to either event. For the sake of simulation, we will say that we have not added both Block-H tags to their respective events. The same would apply for the CMCF categories *Combined Operational Framework* and *MITRE.*

Section 4: Adding MalTech, SecTool, and CWE Tags to the Event

MalTech, although called out in the title of the section, is not a relevant CMCF category for Fault artifacts. Typically, the categories of MalTech and Fault converge when there is information relating to specific a specific SecTool or vulnerability within a MalTech artifact, not the other way around. For example, it is not uncommon for a SecTool like MimiKatz to be mentioned in the body of textual data within a MalTech artifact. MimiKatz is an open-source tool that can be deployed with either malicious or benign intent. For example, a penetration testing team may use MimiKatz as a part of their technical kit during an auditing session for an organization. Therefore, we can pass on this section of the CMCF. Additionally, we had already placed the proper SecTool tags to the events during the event creation process when we added the initial tags to the event.

However, what we must add to both the vulnerability and exploit events are the relevant tags that describe the CWE information. First, we will examine the vulnerability artifact and determine the appropriate CWE. We can see in the title of the artifact that the vulnerability is of an information disclosure type. Within vulnerability linguistical terms, an information disclosure type vulnerability translates to an information exposure weakness within the CWE framework. With this knowledge, we can see that *CWE-200 Information Exposure, CWE-209 Information Exposure Through an Error Message,* and *CWE-538 File and Directory Information Exposure* may apply to the event. However, further review of the vulnerability artifact reveals that the information disclosure vulnerability is not the result of an improperly handled error message.

Additionally, the CMCF does not contain *CWE-200 Information Exposure* within its mapping. Therefore, we cannot attach it to the event. What we are left with is *CWE-538 File and Directory Information Exposure*. Though CWE-200 would have been a natural choice, CWE-538 describes the vulnerability much better because the vulnerability details accurately describe weaknesses within the filesystem.

The exploit artifact is also describing a type of information disclosure. Luckily, with this artifact, we are given the appropriate CWE information by the author of the exploit. Unfortunately, *CWE-532: Inclusion of Sensitive Information in Log Files* is not contained within the CMCF mapping. However, *CWE-538 File and Directory Information Exposure* is. Thus, we

can easily attach CWE-538 to the event with little concern regarding bias or accuracy of recording for this tag. Finally, after we have attached all relevant tags, CVE/CWE information, and textual information to each event, we can publish and close out the process.

Chapter Summary

In this chapter, we went over a methodology for recording SecTool, vulnerability, and exploit OSINT artifacts. We also discussed how these types of artifacts differ from the other artifact types (e.g., Data Breach and MalTech). We also discussed some of the challenges with recording these types of artifacts and the application of CMCF galaxy tags.

Chapter Acronyms

CMCF: Comprehensive Modular Cybersecurity Framework

CVE: Common Vulnerabilities and Exposures

HITRUST: Health Information Trust Alliance

IDPE: Network Intrusion Detection and Prevention Engine (IDPE)

MISP: Malware Information Sharing Platform

Chapter References

Bugtraq: CVE-2020-0728: Windows Modules Installer Service Information Disclosure Vulnerability. (n.d.). Retrieved February 19, 2020, from https://seclists.org/bugtraq/2020/Feb/21

byteGoblin. (2020, February 19). *Nanometrics Centaur 4.3.23—Unauthenticated Remote Memory Leak*. Exploit Database. https://www.exploit-db.com/exploits/48098

Common Weakness Enumeration. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Weakness_Enumeration&oldid=918897150

CVE. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=CVE&oldid=883785290

HITRUST. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=HITRUST&oldid=938925094

MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing (formely known as Malware Information Sharing Platform). (n.d.). Retrieved February 4, 2020, from https://www.misp-project.org/

Suricata IDPE 5.0.2 ≈ Packet Storm. (n.d.). Retrieved February 19, 2020, from https://packetstormsecurity.com/files/156336/suricata-5.0.2.tar.gz

Part 2 – Research

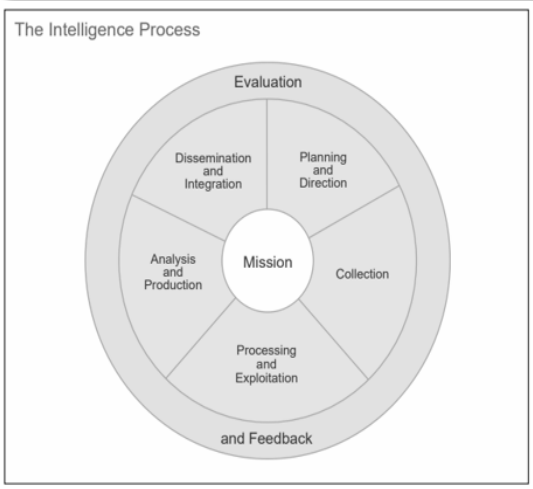
Chapter 5: Research Design and Research Methods

Chapter 5 Introduction

Before we go any further, we need to discuss research, what research is, what research can be, and why it is essential. Intelligence analysts, especially in the cyber domain, are assaulted with a dizzying array of questions from near-peers, tactical teams, risk managers, leadership, and just about any other person interested in information regarding cyber. Contemporary cyber intelligence roles typically revolve around the construct of threat. What is the danger? What is the likelihood of being attacked by the threat? What is the possible impact? These types of questions are a minute sample of the kinds of questions posed to cyber intelligence analysts. Therefore, an analyst must be prepared or educated enough to give as complete an answer as possible when asked the question.

Research helps to aid the analyst in enriching their already overflowing supply of information; or, conversely, develop new knowledge that does not exist. Research is not *threat feeds*. Analysis can be supplemented by threat feed information, but research is not threat feeds. Though relevant to cybersecurity, threat feeds, as a basic description of the phenomena, are automated streams of tactical threat information. These streams can be textual, contain Indicators of Compromise (IoCs), convey threat warnings, or any other type of information. There is no all-in-one format or purpose for threat feeds. However, with research methods, threat feed information can be measured.

If research is not threat feeds, nor casual observation, a collection of information, annotation, or simple essay writing, then what is research? In simple terms, research is a process whereby information is collected, analyzed, and interpreted to elevate an understanding of a particular phenomenon (Leedy & Ormrod, 2015). By this interpretation of research, there is a structured process involved. There is an instrument used to measure a phenomenon. There are tools used to perform an analysis of the outputs of the instrument. Lastly, there is a subject or construct under study. Consequently, as it relates to cyber intelligence, research is the collection, analysis, and interpretation of data that has been derived by some instrument used to measure the phenomenon of cyber.



(JP 2-0, Joint Intelligence, 2013, I-6)

Interestingly enough, a research lifecycle is very similar to what other intelligence analysts from different intelligence domains do in their daily work. An intelligence analyst will collect, analyze, and disseminate data. Researchers will collect, analyze, and disseminate data. So, what is the difference? Frankly, not much. Both the intelligence lifecycle and research lifecycle are performed in a structured format. They both collect data using an instrument, and they both aim to achieve insights from data. Lastly, they both end with an output to an audience. The language used by researchers, as well as intelligence analysts comingles synonymously. Where a researcher aims

to develop results and discussion sessions of a research product, the intelligence analyst will use the terminology of *conclusion and recommendation.* So, in this sense, research and intelligence are the same things, with the most apparent difference in the way language is used by either discipline.

This chapter is primarily based on the book *Practical Research: Planning and Design 11th Edition* (Leedy & Ormrod, 2015). The goal of this chapter is to give an overview of research, research design, and the different research methods that can be applied to the intelligence domain. First, we look at research and design. This first part of the chapter goes over why the design is essential and the various sections of a design that need to be considered before actually conducting the research. The second part of the chapter looks at the different qualitative, quantitative, experimental, historical, and mixed-methods that apply to the collected data, and what the goal of the methods are. This chapter does not go into the interpretation of output from these methods. Nor does this chapter explain the development of a research proposal or the purpose of a literature review. Instead, the aim is to begin to build a lexicon of terminology that helps develop a level-set of terms that can be used by intelligence analysts in peer-review and discussion. Finally, this chapter is not an in-depth review of methodology, analysis, and interpretation.

(Leedy & Ormrod, 2015, p.21)

Research and Design

As with the lifecycle for intelligence analysis, there is a lifecycle for research. Each phase of the lifecycle has a purposeful use within the lifecycle that either builds into or comingles with a corresponding neighbor. Here is the right place within the chapter to review this lifecycle and understand where, when, and why each phase exists.

Phase 1: The Problem

Before any research can begin, there must be an identified problem (or a question). However, the term *problem* here is not defined as something wrong. Instead, the term *problem* is used to describe a gap in understanding. For example, a researcher who studies a drug interaction might merely be looking for information that explains the relationship between age and a particular reaction to the drug. This reaction may or may not be a problem, but it is a part of the observed phenomenon *of*

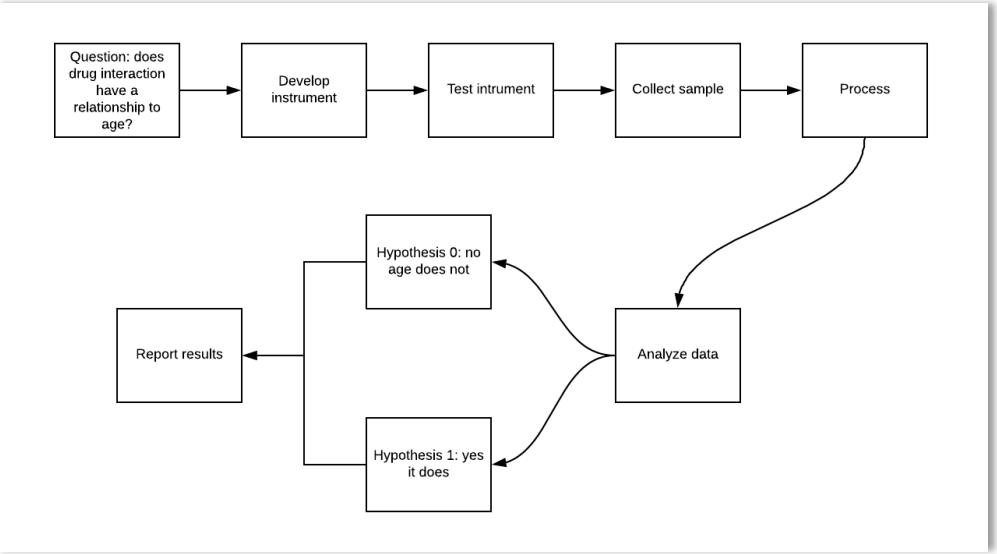
a drug interaction. Sometimes this phase of the research lifecycle is referred to as *developing the problem statement*. We use this problem statement to guide the rest of the research design. The development of a complete problem statement is crucial due to the cascading effect it has on the decision-making process as a researcher proceeds through the rest of the research lifecycle.

Phase 2: Research Goal

During this phase of the research lifecycle, the goal of the research is defined. Typically, a statement about the importance of the study is also laid out during this phase. Setting the purpose of the study helps to reinforce the validity of the problem statement as well as help to assist in determining the actual analysis methods used later on during the analysis portion of the lifecycle. Continuing with the drug interaction, as an example, the goal of the study is to understand if there is a relationship between drug interaction and age. This goal can be measured quantitatively because age is a type of quantitative variable. However, a qualitative research method could be used to help understand the complicated relationship between drug interaction and age; it makes more sense to use a quantitative method to achieve the goal of understanding this interaction more concisely.

Phase 3: Research Architecture

A design architecture is a high-level perspective of the overall researcher design. It lays out strategic requirements, begins to develop the subsections of design, and helps to lay out the foundations of a hypothesis(es) to be tested. Sometimes it helps to represent an architecture visually. However, that is not a requirement for the research lifecycle to be successful. The following is an example of the drug interaction research lifecycle drawn out as a generalized architecture.



The research question influences the research design architecture, moves to the development of an instrument based on the research question and research methodology, which, in this case, we previously identified as a quantitative methodology. We then test the instrument for validity as well as reliability, collect the sample data, process (or store) the data, analyze the data using a methodology, interpret the results, accept or reject a hypothesis, and finally report the results. Each part of this architecture can be identified for possible strengths and weaknesses as they align to the problem statement as well as the goal of the research. Additional pitfalls can be identified during the testing of the instrument. Lastly, the architecture allows the researcher to develop more granular goals per item of the architecture. For example, the development of the instrument can be broken down into sub-tasks or sub-questions. As an example, a researcher might ask questions like which software is best for this type of measurement, how long would it take to configure the software to the study, and is the software easily accessible to the researcher?

Phase 4: Hypothesis Development

In terms of research, a hypothesis is generally something that a researcher states as a possible assumption regarding the results of a study. A research endeavor can have more than one hypothesis, but will always have a *null hypothesis*, or a statement regarding the observance of no change in information after the data has been analyzed and interpreted. A hypothesis is also based on a sample of a population rather than an entire population. Therefore, due to statistical chance, can never be absolutely true. Instead, a researcher either accepts or rejects a set of findings and their relationship to a hypothesis. Once more, using the drug interaction example, a *null hypothesis* may be that there is no relationship between drug interaction and age. Depending on the interpretation of the data analysis, a researcher may or may not accept the null hypothesis. The language used in the hypothesis is just as important as the language used to develop the problem statement. Inconsistency alters the relationship between the research problem and hypothesis; and, fundamentally invalidates any of the findings during the analysis of the data.

Phase 5: Research Development and Design

A research design is the more granular and precise translation of the operational portion from the research architecture. During the design phase, a researcher will determine things like the sample, the sample demographics, the exact variables, where the variables reside in relationship to each other (e.g., independent, dependent, moderating et al.), the frequency of sample measurement, the exact tools for collection, how data will be stored, correct analysis methods, and any other granular item required to complete the research.

Phase 6: Data Analysis

The data analysis phase of the research lifecycle is a post sample collection function that applies the selected analysis method to the data that was collected on the sample. It is essential to understand that data analysis is not an interpretation of the results. It is only the function of applying analytical methods to the data. The selection of analysis methods, as previously mentioned, is determined by the problem statement, which then drives the rest of the research lifecycle. In the drug interaction example, we decided that the analysis method would be quantitative. For this, we could look at descriptive statistical methods like frequency analysis, or we may want to develop a more rigid statistical result. Therefore, we might use a Paired Sample t Test to show whether or not the data is statistically significant.

Phase 7: Interpretation and Results

Following the data analysis phase of the research lifecycle is the interpretation and results phase. The interpretation of results is one of the last steps in the research lifecycle. It is during this phase where a researcher would examine the output of the analysis and make their findings. These findings are typically conveyed in a result-based discussion about the interpretation of the output, the limitations of research, how these limitations affected the study, and possible new avenues for further investigation of the phenomena. The interpretation and results phase is also where the researcher would either accept or reject the hypothesis. The acceptance or rejection is the researcher's conclusion opinion of the output backed by the analysis method.

Research Methods

Samples and Variables

Samples

Research is typically conducted on a sample. The sample is representative of a population. The sample is not the entire population. Instead, samples are purposeful representations of the population based on population attributes or demographics. In our drug interaction example, we did not define the sample, but let's say we had. What if the sample only contained males or only individuals of a particular nationality? This demographic determination would affect the interpretation of the output from the analysis. We know from the example that we are looking to see if there is a relationship between drug interaction and age. However, beyond that, we cannot say that gender had a moderating or mediating effect when interpreting the output because it was not included in the problem statement; only that there was a curiosity to determine if there was some relationship between drug interaction and age. Therefore, when producing the results, as a researcher, we would have to express this limitation as a part of the research design. We may, at

that point, and during a results discussion, include further study recommendations that investigate if there is a difference when gender is added into the design.

Variables

Before discussing the parent categories of methodologies, we must first look at how the data is classified into variables. In research, variables have classifications that affect which methods they can be used in, as well as what instruments can be selected to measure the variable. There are four general classifications that data can be a part of: nominal, ordinal, interval, and ratio. Since research methodologies somewhat differ in terms of mathematics (e.g., algebra vs. statistics), with statistics being the predominant mathematical framework for analysis, the classifications allow for the measurement of different data types. For example, the word *male* is different than the value *100*. However, both can be classified into variables, measured, and compared with statistics.

Nominal Variables

Nominal variables are categorical or strings that represent a group. Using the drug interaction example, we have age as a nominal variable, but let's say that we wanted to know if there was a difference between genders. The gender of an individual now becomes a variable of the study, and because it is categorical, we would classify gender as an additional nominal variable of the study. Other nominal variable examples would be race, religion, and eye color.

Ordinal Variables

Ordinal variables are those that are *greater than* or *less than* one another. They are like nominal variables in the fact that they are somewhat categorical, but unlike nominal variables, they have inferred weights between them. Another way to consider ordinal variables is that an ordinal variable is like a ranking system. An example of an ordinal scale would be to rank something with low, medium, high, and very high context. A ranking of low is lower than medium, so on and so forth.

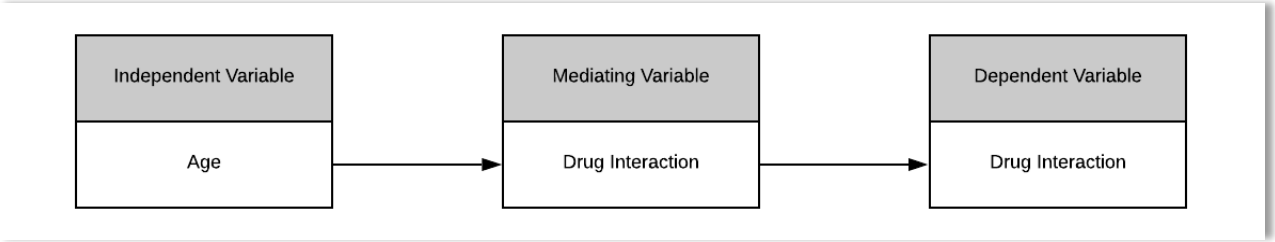
Interval and Ratio Variables

Interval variables are those that have an arbitrary zero point and equal units of measurement. Interval variable values of zero do not infer that something is absent. Instead, zero in an interval variable scale means that it is the absolute lowest point on the scale. Temperature is often used as an example of an interval variable scale (Leedy & Ormrod, 2015). A temperature of 0^{o} Fahrenheit does mean the absence of heat. Conversely, ratio variables, which are very similar to interval variables equal units of measurement, define the zero point of a scale as the absence of an observable.

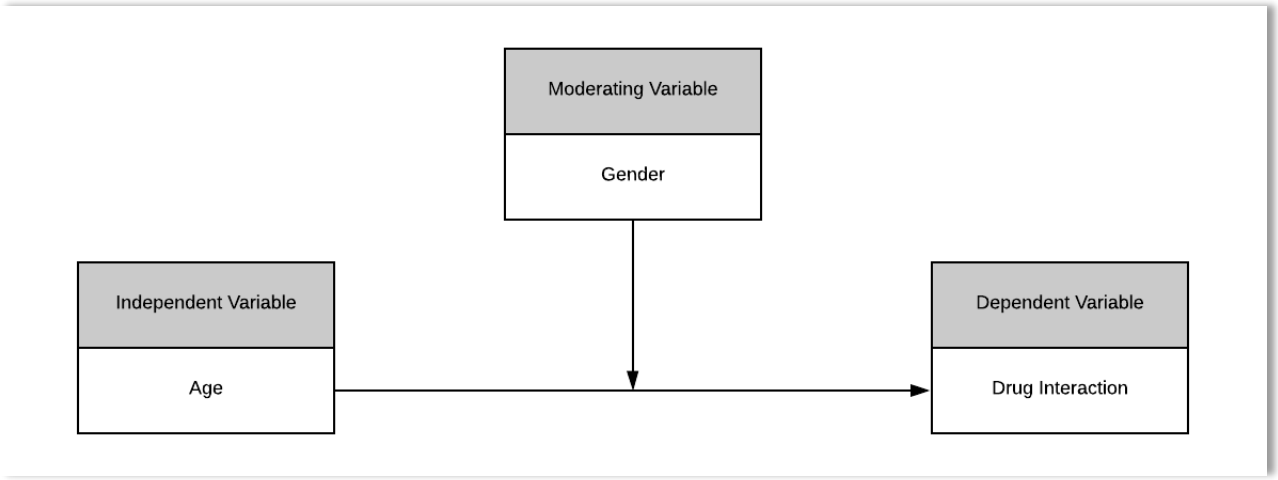
Variable Identity

Cause and effect. These are the two primary tenants of *variable identity.* Not only do variables that are used in research have classifications, but they also have an identity that describes their role in the relationship. There are four primary identities for variables: independent, dependent, mediating, and moderating. Each has a position within the relationship between variables that must be identified by the researcher for the analysis to make sense. Again, using drug interaction as an example, the dependent variable is *drug interaction,* and the independent variable is age. The identity of the variable helps to understand its role in causality.

Mediating variables support the explanation of causality, or why a particular independent variable affects a dependent variable. The mediating variable is also a variable that is influenced by the independent variable. Continuing with the drug interaction example, a mediating variable in this particular study might be the amount of drug that is administered if a person is a specific age. The amount of drug is influenced by age, but the amount also plays a part in the drug interaction outcome.



The moderating variable, like the mediating variable, affects the outcome of the analysis. The difference between a moderating variable and a mediating variable is that the mediating variable is not influenced by the independent variable but does influence the causality. Again, using the drug interaction example, let's use gender as a mediating variable. Gender is not influenced by age, but we know that there is a difference between male and female physiology. Hence, gender does affect drug interaction.



Types of Error

There is always a chance for error when testing a single or set of hypotheses. These types of probabilities for error are categorized as *Type I* and *Type II* error. The two different types of errors are introduced when testing for statistical significance. Probable error values are compared against the *alpha* value, which is predetermined before the analysis phase of the research lifecycle. Typical *alpha* values are range between .05 and .01. The value of alpha directly impacts acceptance of the null hypothesis, which, in turn, increases or decreases the chance of *Type I* or *Type II* error. For example, lowering the *alpha* value below .05 reduces the chance of *Type II* error, but increases the chance for *Type I* error. Determining the *alpha* value is a game of cat-and-mouse when dealing with error in hypothesis testing.

Type I Error

A Type I error happens when a researcher rejects a correct null hypothesis.

Type II Error

A Type II error happens when a researcher accepts a false null hypothesis.

Qualitative Methodology

One of the parent categories of methodologies that a researcher can use as a research design is the *qualitative methodology.* Broadly speaking, qualitative research methodologies seek to understand the phenomena in its natural settings instead of placing controls like in an *experimental* method. Furthermore, the qualitative approach does not attempt to reduce the

complexity of a phenomenon. What's more, is that qualitative methodology does not seek to identify a causality from a phenomenon.

Case Study

Case studies are a study of a particular individual, group, or event for a period of time. The various types of instruments used to collect data include observational notes, documents, interviews, video, and previous records like scores from a test. Case studies also require the researcher to spend extensive amounts of time with the individual or within the environment being studied. The primary goal of a case study is to collect enough data to sufficiently describe the phenomena in a full context.

Ethnography

An ethnographic study is like a case study in that it studies and individual, group, or event for a period of time. However, the difference is that unlike a case study that aims to provide a full context of a phenomenon, the goal of an ethnographic study is to portray the depth of the phenomenon. For example, a sociologist may conduct ethnographic research on a particular tribe to determine their daily habits, colloquialisms, and cultural beliefs. Interviews, observer notes, and the recording of audio and video data can be used as collection mechanisms during an ethnographic study.

Phenomenological Study

A qualitative research design that includes a phenomenological study as its methodology has a goal of gaining insight into how the individual or group perceives the phenomena. Unlike an ethnographic study, the phenomenological research typically requires a careful sampling of a population within the phenomena and the use of extensive interviews. Let's say instead of measure drug interaction and age; we wanted to understand what the individual experienced when they took the drug. Some questions that might get answered during this type of study is how the drug made the participant feel, did they notice any changes, and what were their social interactions like after taking the medication.

Grounded Theory Study

Grounded theory studies begin with data and then attempt to create a theory based on the data. Like the case study, ethnographic study, and phenomenological study, the researcher will spend time with the individual, group, or event to collect data. The primary procedural difference is that a grounded theory study will already have developed a particular variable set for which to begin measurement immediately.

Content Analysis

The content analysis method uses a systematic process to examine a body of material to identify any patterns, thematic iterations, and bias within the material. Documents, journals, books, and transcripts are just a few examples of the types of medium that would be used in a content

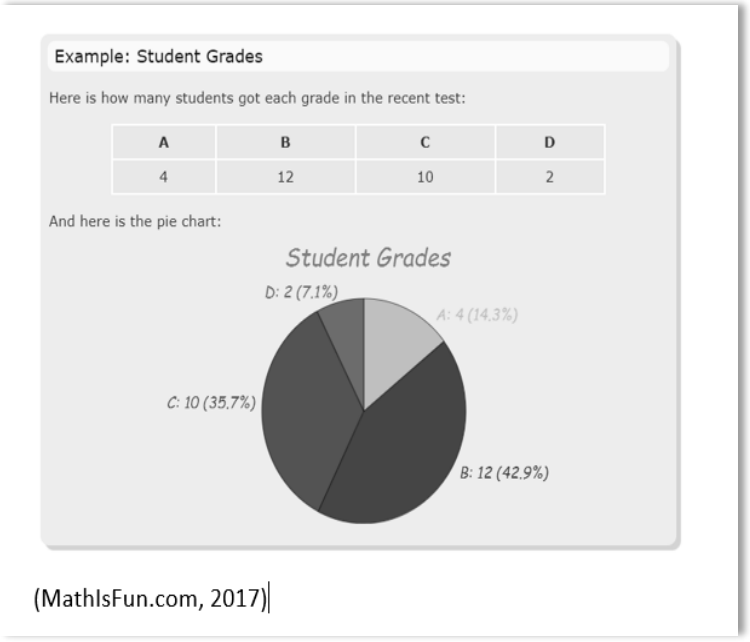
analysis methodology. This particular methodology is also used as a complementary methodology conjoined into more extensive research design to support the research architecture.

Quantitative Methodology

Research that uses quantitative methods in their design typically seeks to answer problem statements in terms of causality. This is not to say that the other research methods (qualitative, experimental, or historical) do not have quantitative components. Quantitative studies do not have to rely on people as samples to determine causality. Simply, quantitative methods of research can be sufficiently applied to living and non-living phenomena to study causality. Additionally, quantitative studies are geared heavily for the use of statistical mathematics as a way of describing causality.

Descriptive Statistics and Central Tendency

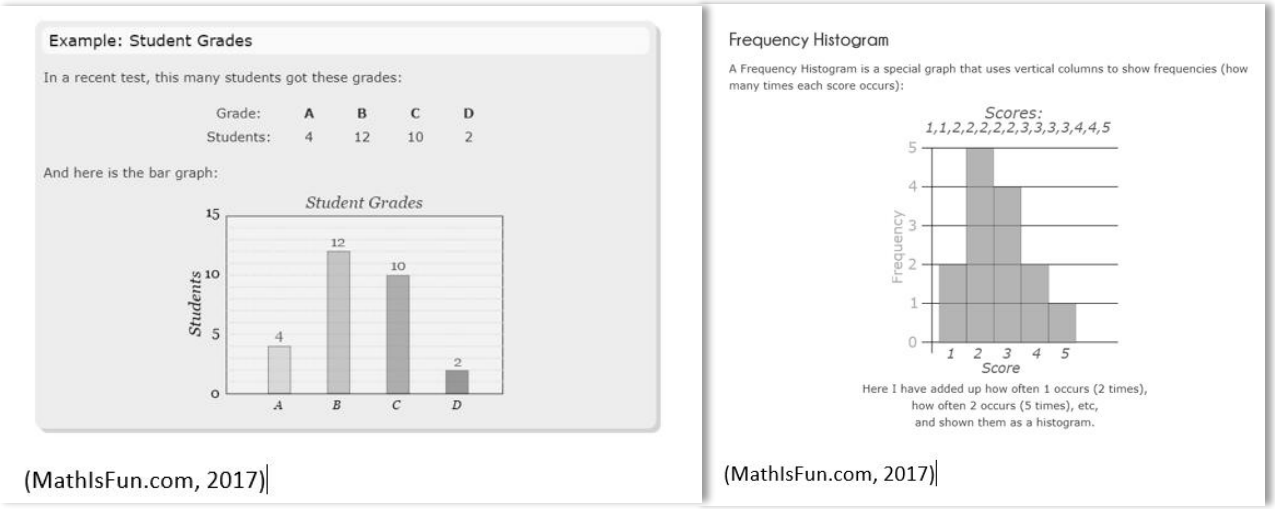
Many of the statistical methods that can be applied to data sets look to examine the difference between means (or averages). These mathematical methods also seek to understand the central tendencies within the data. Points of central tendency include the *mean (or average), mode,* and *median* of a data set. Though considered mathematically trivial, these foundational points of central tendency are the starting points for a more in-depth mathematical examination of the data.



(MathIsFun.com, 2017)

Additionally, these points of central tendency are found when their mathematical formulae are applied to interval and ratio data. The *mean* is the average of a data set. It is the total of the data divided by the number of data points. The *median* is the center of the data set. It is the resultant number that sits in the middle of the range. Finally, the *mode* of a data set is the resultant number that occurs most frequently in the data set. Some ways to represent descriptive statistics are with visualizations like a bar chart, pie charts, and histograms.

Example Bar Chart and Histogram



Statistical Methods for Quantitative Research

Several quantitative statistical methods can be used in quantitative research designs. The caveat is that each technique holds assumptions about the variables being analyzed. These assumptions are why it was essential to introduce the different variable types before discussing specific methods, premises, and outputs. Furthermore, each statistical method has an interpretation to the result that is unique to the particular method. Though many methods reveal something about the statistical significance, which, in turn, shows the possible types of error, how the output is determined is unique. Now is also an excellent time to discuss statistical significance from the viewpoint of how the term *significant* is used interchangeably during the research lifecycle. Let's say we are studying the effect a rule tuning may have on the number of false positives reported by a firewall using a pre/post analysis of the data. Our hypothesis is as follows:

$H_{1\ (null\ hypothesis)}$: There is no change in the number of false positives after rule tuning

$H_{2\ (alternative\ hypothesis)}$: There is a change in the number of false positives after rule tuning

Here we are comparing the means of two independent data sets before and after a rule tuning has been implemented. The null hypothesis in this study states that there is no change in the number of false positives after rule tuning. For this, we can use a *Paired t-test* to compare the two groups to each other. A Paired t-test assumes that you have a single nominal variable. In this case, we are using the term *False Positive* as the nominal variable, and the frequency of false positives as the second variable that we have taken under two different conditions (e.g., pre/post). Our *alpha* value for this test is set to .05. Therefore, to be *statistically significant,* we would need to see an output from the analysis at or below .05 to be considered *statistically significant* to a point whereby we would reject the null hypothesis and say that there is a change in the number of false positives. However, let's say our output significance value is above .05. This larger value would mean that there is no *statistical significance,* and therefore we would accept the null hypothesis. Yet, this does not mean that there is not a significant reduction in

false positives that would be considered a worthwhile effort from the rule tuning. If the average number of false positives were reduced by 10% of the typical central tendency, then from a decision support viewpoint, we might say that the rule tuning effort was a success. So, though not *statistically significant*, the study showed the attempt to be meaningful to the group who might have to manage the number of false positives daily.

The following are models of specific statistical methodologies, their assumptions, and the interpretation of outputs. This information has been taken and synthesized from the Kent State University Libraries tutorials for statistical methods, and a complete citation has been included at the end of the chapter. These examples also assume that the reader has access to statistical software like the Statistical Package for the Social Sciences (SPSS)[2].

Descriptives

			Statistic	Std. Error
Height	Mean		68.0318	.26366
	95% Confidence Interval for Mean	Lower Bound	67.5135	
		Upper Bound	68.5501	
	5% Trimmed Mean		67.9687	
	Median		67.5700	
	Variance		28.363	
	Std. Deviation		5.32566	
	Minimum		55.00	
	Maximum		84.41	
	Range		29.41	
	Interquartile Range		6.78	
	Skewness		.230	.121
	Kurtosis		.113	.241
Weight	Mean		181.0316	2.20465
	95% Confidence Interval for Mean	Lower Bound	176.6966	
		Upper Bound	185.3666	
	5% Trimmed Mean		178.4763	
	Median		172.9600	
	Variance		1827.535	
	Std. Deviation		42.74968	
	Minimum		101.71	
	Maximum		350.07	
	Range		248.36	
	Interquartile Range		50.62	
	Skewness		1.005	.126
	Kurtosis		1.502	.251

(Kent State University Libraries, 2020)

Descriptive Statistics for One Numeric Variable – Exploratory

This methodology is used to understand through exploratory function a single univariate set of data. What this means is that we are merely looking exploratorily at how a unique collection of data lives within its central tendencies.

Assumptions:

It is assumed that we have a single numeric set of data for a single variable.

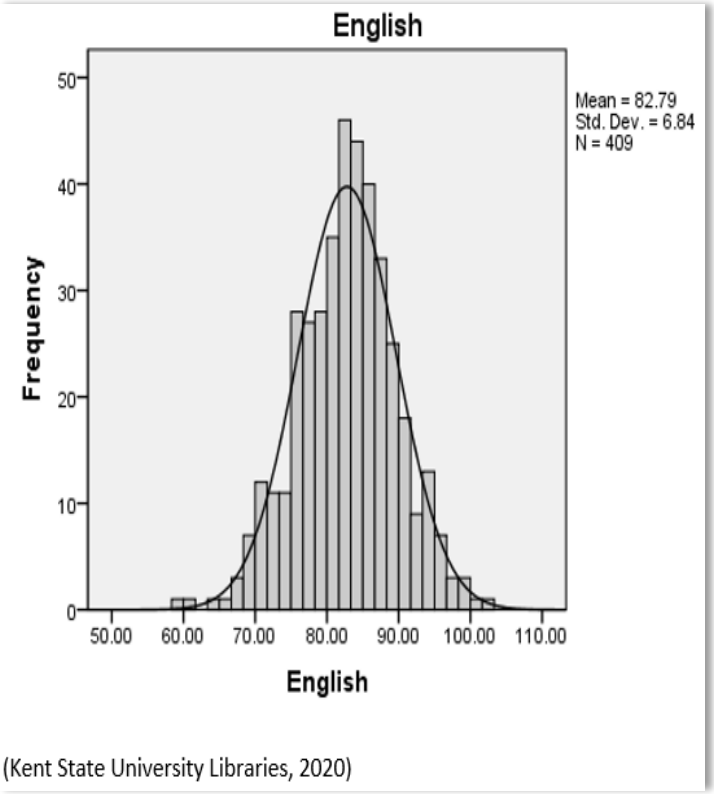
It is assumed that we have a single set of categorical variables all classified under a parent within a categorical hierarchy.

Outputs:

List of descriptive statistics (number of cases, mean, median, variance, standard deviation, minimum, maximum, and range).

[2] (IBM SPSS Statistics—Overview, 2020).

Descriptive Statistics for One Numeric Variable – Frequency



(Kent State University Libraries, 2020)

Frequencies are useful for determining the center of the data, how the data is spread out over the range, what the extreme values may be, and the overall shape of the distribution.

Assumptions:

The variables are continuous, interval, or of the ratio type.

There is at least one set of data for these types of variables.

Outputs:

Typical outputs from an analysis of frequency include, but are not limited to, mean, standard deviation, variance, sum, minimum, maximum, range, and percentiles. These outputs can be represented with bar charts, pie charts, and histograms.

Descriptive Statistics for Many Numeric Variables – Descriptive

We can also perform descriptive statistical analyses on more than one variable at a time. The methodology is the same, and the output reports are relatively similar to those applied to a single variable. The assumptions are also the same when applying this method to more than one variable. Below is an example output showing the descriptive statistics for more than one variable.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
English	431	59.83	101.95	82.7265	6.82982
Reading	435	55.11	103.62	82.0394	7.63745
Math	435	35.32	93.78	65.4512	8.29165
Writing	435	64.06	93.01	79.5392	5.50151
Valid N (listwise)	431				

(Kent State University Libraries, 2020)

Descriptive Statistics by Group – Comparing Means

Comparing means is used for examining the differences in means between groups within a parent level category. For example, using the false positive example, let's say we wanted to compare the points of central tendency between false positive and true positive events. These events are sub-categories to the parent hierarchy *Event Type*. Under each sub-category, there is a count of occurrences that allows us to use these occurrences as the ratio variable type. Our results from performing statistical analysis that compares means will show us the differences in central tendency between the two groups.

Assumptions:

It is assumed there is a parent category of a nominal variable with at least two levels.

It is assumed that there is at least an interval or ratio data for each level and is independent of each other.

It is assumed that there is at least a dependent variable and an independent variable.

Outputs:

There are several outputs from this statistical analysis procedure that includes, but are not limited to, mean, standard deviation, range, sum, and variance.

Report

Mile time

Are you an athlete?	Gender	Mean	N	Std. Deviation	Minimum	Maximum
Non-athlete	Male	0:07:48	91	0:01:36.116	0:05:05	0:12:21
	Female	0:10:06	130	0:01:41.347	0:06:56	0:14:02
	Total	0:09:09	221	0:02:00.024	0:05:05	0:14:02
Athlete	Male	0:06:46	90	0:00:48.962	0:05:03	0:08:56
	Female	0:07:00	72	0:00:48.770	0:05:16	0:08:41
	Total	0:06:52	162	0:00:49.253	0:05:03	0:08:56
Total	Male	0:07:17	181	0:01:22.439	0:05:03	0:12:21
	Female	0:09:00	202	0:02:04.103	0:05:16	0:14:02
	Total	0:08:11	383	0:01:58.036	0:05:03	0:14:02

(Kent State University Libraries, 2020)

Frequency Tables

A frequency table is very similar to the comparison of means. However, frequency tables do not require dependent and independent variables to be known. Instead, a frequency table is a summary of descriptive frequencies about a set of variables (nominal, ordinal, interval, or ratio). The assumption is that there is more than one variable that requires frequency analysis. The output is very similar to the comparison of means and can be placed into a pie, bar, and histographic charts for visualization.

A frequency table will typically include four columns to represent the results: the frequency, percent, valid percent, and cumulative percent. The frequency column represents the total number of observations. The percent column indicates the percentage of observation by category. Valid percent only reports results from non-missing cases, and the cumulative percent column is for the displaying of output percent for the sample.

Class rank

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Freshman	147	33.8	36.2	36.2
	Sophomore	96	22.1	23.6	59.9
	Junior	98	22.5	24.1	84.0
	Senior	65	14.9	16.0	100.0
	Total	406	93.3	100.0	
Missing	System	29	6.7		
Total		435	100.0		

(Kent State University Libraries, 2020)

Crosstabs

Class rank * Do you live on campus? Crosstabulation

Count

		Do you live on campus?		
		Off-campus	On-campus	Total
Class rank	Freshman	37	100	137
	Sophomore	42	48	90
	Junior	90	8	98
	Senior	62	1	63
Total		231	157	388

(Kent State University Libraries, 2020)

Class Rank * Do you live on campus? Crosstabulation

			Do you live on campus?		
			Off-campus	On-campus	Total
Class Rank	Underclassman	Count	79	148	227
		% within Class Rank	34.8%	65.2%	100.0%
	Upperclassman	Count	152	9	161
		% within Class Rank	94.4%	5.6%	100.0%
Total		Count	231	157	388
		% within Class Rank	59.5%	40.5%	100.0%

(Kent State University Libraries, 2020)

Unlike frequency tables that only look at describing the frequencies for a single categorical group, Cross-tabulation (a.k.a. crosstabs) is a way to describe the frequencies between two categorical groups. The way this is represented is by row and column designation by grouping. One group creates the rows, while the other creates the columns of the table. Percentages are typically based on a specific view requested of the data. However, percentage results are not a requirement of cross-tabulation.

Chi-Square Test of Independence

The Chi-Square Test of Independence, also known as the Chi-Square Test of Association, is a non-parametric test that determines a level of independence or association between categorical variables. A Chi-Square test can also be used to determine if there was statistical significance found within the analysis of the data.

Do you smoke cigarettes? * Gender Crosstabulation

Count

		Gender		
		Male	Female	Total
Do you smoke cigarettes?	Nonsmoker	149	148	297
	Past smoker	13	24	37
	Current smoker	31	37	68
Total		193	209	402

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	3.171[a]	2	.205
Likelihood Ratio	3.217	2	.200
Linear-by-Linear Association	1.106	1	.293
N of Valid Cases	402		

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 17.76.

(Kent State University Libraries, 2020)

Assumptions:

There are two categorical variables, and each of these variables has at least two levels.

Each category of a variable is independent of each other.

The expected frequencies per cell are above a numerical value of 1.

At least 80% of the cells examined have a numerical value greater than 5.

Hypothesis testing looks to accept or reject an association between variables.

Variable examples:

Event Type, which is then broken down into false positive or true positive.

Time of Day which is broken down to first half (00:00 to 11:59) and second half (12:00 to 23:59)

Outputs:

The output from a Chi-Square Test of Independence starts with a cross-tabulation performed on the two categorical variables. The next portion of the process that produces the results of the Chi-Square Test of Independence is the actual mathematical function of the Chi-Square Test of Independence, which uses the data from the cross-tabulation table.

A p-value is calculated during the Chi-Square Test of Independence.

The p-value is used to accept or reject the hypothesis of association. A p-value equal to or less than the set alpha value for the study results in the rejection of the null hypothesis and the acceptance of the alternative hypothesis. A p-value higher than then alpha value set for the study results in the opposite acceptance and rejection of the null hypothesis.

Person Correlation

This statistical method produces what is called a correlation coefficient, or *r*. This coefficient is used to measure the direction and strength of a linear relationship. Some of the common uses for this particular statistical method would include studies looking to determine the correlation

between pairs of variables, or correlations between (as well as within) sets of variables. The Pearson Correlation, like the Chi-Square Test of Independence, produces a significance value (or sig. value) that can be used to determine statistical significance.

Assumptions:

The data being studied is from two or more continuous variables (e.g., ratio or interval).

There is an assumption of a linear relationship between the variables.

The values for all variables are unrelated.

Variables, in one case, cannot influence the value of variables in another case.

Variables are normally distributed.

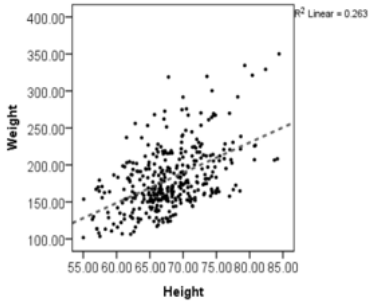
The samples are randomly generated, and there are no outliers.

Outputs:

Results of the correlation coefficient range between -1, 0, and 1.

The closer to zero the correlation coefficient, the lesser the strength of the relationship between variables.

Results can be portrayed with a scatter plot that includes a line of best fit to represent the direction of the linear relationship.



Correlations

		Height	Weight
Height	Pearson Correlation	1	.513**
	Sig. (2-tailed)		.000
	N	(A) 408	(B) 354
Weight	Pearson Correlation	.513**	1
	Sig. (2-tailed)	.000	
	N	(C) 354	(D) 376

**. Correlation is significant at the 0.01 level (2-tailed).

(Kent State University Libraries, 2020)

One Sample t Test

One-Sample Test

	Test Value = 66.5 (A)					
	(B) t	(C) df	(D) Sig. (2-tailed)	(E) Mean Difference	95% Confidence Interval of the Difference (F) Lower	Upper
Height	5.810	407	.000	1.53176	1.0135	2.0501

(Kent State University Libraries, 2020)

This type of statistical method looks to determine statistical significance between a single variable mean and a hypothesized population mean. This comparison between means assumes that the One Sample t Test can only be used to compare a sample means to a specific constant value mean.

Assumptions:

The test variable is continuous (e.g., ratio or interval).

Variable scores are independent of each other and do not influence values between cases.

Sample data is a random sampling from the population.

The variable data is normally distributed.

There are no outliers.

Outputs:

The outputs from a One Sample t Test include but are not limited to the t Statistic, degrees of freedom, and significance (or sig.). The significance value can be used to determine if the null hypothesis is accepted or rejected.

Paired Sample t Test

This type of t Test compares means between two sets of data. Some of the common uses for a Paired Sample t Test included but are not limited to measurements taken at two different time intervals, measures taken from two halves of a single subject, and measurements taken under two different conditions.

Assumptions:

The dependent variable being tested is continuous (e.g., interval or ratio).

Subjects in the first group are the same in the second group.

The data is normally distributed.

There are no outliers in either group.

Outputs:

The outputs for the Paired Sample t Test are similar to the One Sample t Test in that they both report results that can be used to determine statistical significance; and, whether or not to accept the null hypothesis.

Paired Samples Test

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
English - Math	17.30	9.50303	.4763	16.3608	18.2337	36.313	397	.000

(Kent State University Libraries, 2020)

Independent Samples t Test

This particular statistical method examines the difference in means between two independent groups and looks for statistical evidence of an association between population means. The Independent Samples t Test is commonly used when looking at the statistical differences between the means of two groups, two interventions, or two change scores.

Assumptions:

The dependent variables for the study are either interval or ratio.

The independent variable is categorical.

Cases have values in both the independent and dependent variables groups.

The samples are independent of each other, and there is no relationship between samples.

Subjects from the first group cannot be subjects of the second group.

The data is normally distributed.

Outputs:

The Independent Samples t Test reports the results from the analysis in much the same way as the other t Tests in that there is a significance value reported that could be used to determine statistical significance, as well as the acceptance or rejection of the null hypothesis. However, this particular type of t Test also includes what is called *Levene's Test for Equality of Variances*. The hypothesis for the Levene's Test for Equality of Variances is as follows:

$H_{1\ (\text{null hypothesis})}$: the variances of group 1 and 2 are equal

$H_{2\ (\text{alternative hypothesis})}$: the variances of group 1 and 2 are not equal

Interpreting the results of an Independent Samples t Test is a bit confusing. The significance value only takes up one row of the output but does not represent the results for the row for which it is found. The significance found in the results is not only used for accepting or rejecting the null hypothesis but also indicates which portion of the Levene's Test for Equality of Variances is used to describe the results.

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig. (A)	t (B)	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Mile time	Equal variances assumed	102.98	.000	13.475	390	.000	0:02:14	0:00:10	0:01:55	0:02:34
	Equal variances not assumed			15.047	315.846	.000	0:02:14	0:00:08	0:01:57	0:02:32

(Kent State University Libraries, 2020)

One-Way ANOVA

The One-Way Analysis of Variance (ANOVA) is used to compare means between groups when the number of groups exceeds two. The use of a One-Way ANOVA can be found when comparing differences in means between more than two groups, more than two interventions, or more than two change scores. Some research designs that might include a One-Way ANOVA would be Field Studies, Experiments, and Quasi-Experiments.

Assumptions:

The dependent variable is either interval or ratio.

The independent variable is categorical with at least two levels.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	26.788	2	13.394	9.209	.000
Within Groups	509.082	350	1.455		
Total	535.870	352			

(Kent State University Libraries, 2020)

Cases have values on both the dependent and independent variables.

Samples are independent of each other.

Subjects in the first group cannot be in any other group.

Subjects cannot influence the subjects in another group.

The data is normally distributed.

There are no outliers.

Experimental Methodologies

Sometimes research looks to examine the causal nature between variables. Though quantitative methodologies can compare means, develop a hypothesis for testing, and seek to understand the statistical significance, quantitative methods offer no control between observations. Experimental methods, by design, consider the different variables that may have causal effects on the relationship between variables and then institute controls into the design to capture how these influencers affect the relationship.

There are several main types of experimental methodology designs: Pre-Experimental, Experimental, Quasi-Experimental, Ex Post Facto, and Factorial design. A Pre-Experimental design is used when it is not feasibly possible to show causality from the results of the research. The Pre-Experimental design is typically used when formulating a tentative set of hypotheses used in later, more controlled studies. Within the context of true Experimental design, a researcher will manipulate the independent variable to see how it affects the dependent variable. Quasi-Experimental designs emphasize the importance of randomness in either the selection of

the sample or in the introduction of different treatments to the group. In Ex Post Facto designs, the researcher will identify conditions that have already occurred and are still conditionally present within the sample attributes. Lastly, Factorial design for experimental research is used when a researcher is studying two or more independent variables.

Confounding Variables

These types of variables are conditions within the study that are difficult to draw causal conclusions for *why* or *how* the particular variable affects the causality between variables. One way to control for these types of variables is to implement controls that keep the confounding variable at a constant value. Introducing a control group into an experimental design is another way of attempting to control for confounding variables. Random group assignments, pretests for equivalence, and exposure to all introduced treatments are other ways for controlling confounding variables.

Pre-Experimental Design: One-Shot Experimental Case Study

This type of Pre-Experimental design looks to simply observe the results of an introduction of treatment into a group. There is low validity to this type of design since it is almost impossible to determine if the result is actually due to the introduction of the treatment into the group. This low validity means that within the context of this design, there is no way to know if the observable was pre-existing before the introduction of the treatment. However, this type of design is easy to construct and can be used to develop a more mature hypothesis as well as lead to a more stringent research methodology.

Pre-Experimental Design: One-Group Pretest-Posttest Design

In a One-Group Pretest-Posttest design, a group is measured before and after the introduction of treatment. Unlike the One-Shot Experimental Case Study design, in this type of Pre-Experimental design, the researcher has taken measurements before introducing the treatment and therefore knows if the effect exists or not. However, this type of design does not control well for confounding variables.

Experimental Design: Pretest-Posttest Control-Group Design

This type of Experimental Design places randomly places sample assignments into either the experimental group or the control group. The experimental group receives the treatment, while the control group does not. Furthermore, the control group is cordoned off from any influence the experimental group may have on it. This type of design helps to control for confounding variables as well as determine if the treatment affected a change after its introduction into the experimental group.

Experimental Design: Solomon Four-Group Design

The Solomon Four-Group design addresses the effect that pretesting may have on a group. In a Pretest-Posttest Control-Group design, there are only two groups. The addition of two other groups into the design enhances the validity of the results. The drawback is that it requires twice the sample size than if using a Pretest-Posttest Control-Group design.

Quasi-Experimental Design: Nonrandomized Control-Group Pretest-Posttest Design

In a Nonrandomized Control-Group Pretest-Posttest design, there are two groups under examination. However, unlike the Pretest-Posttest Control-Group design, the subjects are not randomly placed into either the experiment or control group. This design does not allow for the determination of similarity during pretesting. However, it does consider pretesting results. Therefore, a researcher can view the results as being a part of the treatment.

Quasi-Experimental Design: Simple Time-Series Design

The Simple Time-Series design takes several measurements of the dependent variable over a set of timed intervals. This type of Quasi-Experimental design only observes the results of treatment from one group. There is no differentiation between an experimental or control group. The measurements taken before the introduction of the treatment are considered the baseline. The lack of a control group in this design allows for confounding variables to be a part of the causality explanation, possibly.

Ex Post Facto Design: Simple Ex Post Facto Design

A research study that includes a Simple Ex Post Facto design is looking to measure the experience rather than the treatment. In this Ex Post Facto design, the researcher is not involved nor responsible for the introduction of the treatment. Instead, the treatment was introduced into the sample long before the researcher began the study. Furthermore, studies that include this type of design are looking only to conclude that certain variables appear to be associated to a pre-existing condition.

Factorial Design: Two-Factor Experimental Design

There are two independent variables and a minimum of four groups under observation within a Two-Factor Experimental design. The results of the effect of the variables on the group are measured between the groups. The Two-Factor Experimental design is very similar to that of the Solomon Four-Group design in that it is comparing several groups to each other to determine the causality from the introduction of treatments.

Historical Methodologies

Research that includes Historical methods as a part of the research design is looking to derive meaning from supposedly random events. From this information, researchers can speculate the causality of events, make inferences about their relationships, and draw conclusions about the effects the events have on those who have participated in the events.

Data collected using Historical methods are separated into two categories. The first is the *primary source*. This source of data is generated chronologically and typically involves events that appeared first in a timeline. Primary sources can come in the form of newspaper articles, laws, census reports, deeds, photographs, films, and paintings. The other category, *secondary source*, is from those who have synthesized the *primary source* data into collections. An excellent example of a *secondary source* is a book published by a historian regarding a particular historical event.

Mixed-Methods

Research isn't meant to constrain the researcher. Frankly, the world of observing phenomenon is a dynamic one. There is no one right way to study something. Sometimes a researcher needs to combine methods from the various domains to make sense of what the researcher is observing. Therefore, a choice of Mixed-Methods is used for studies whose phenomenon falls outside the scope of obvious observation and method. Typically, a Mixed-Method design will include both qualitative and quantitative methods in a single research activity. The quantitative portions of the study are used to complement the qualitative parts in a way that enhances the completeness of the research.

Convergent and Embedded Designs

In a Convergent design, the research collects both the qualitative and quantitative data at the same time. The data is given the same priority or weighting, and the overall focus of the design is the triangulation of the results. This goal of triangulation means that both the qualitative and quantitative data support the same conclusion. However, this type of design does not ensure that the outcome between qualitative and quantitative data does, in fact, come to the same conclusion. Like the Convergent design, the Embedded design follows a similar process for collecting the data in parallel. The difference between Convergent and Embedded is that the Embedded design prioritize one type of data over another. For example, a researcher performing research using an Embedded design favors or prioritizes quantitative over qualitative data.

Exploratory and Explanatory Designs

Both Exploratory and Explanatory research design methods incorporate a two-phased approach to the collection of the data. Additionally, both design types use qualitative and quantitative data. The difference is in the phase for which the type of data is collected and how that data influences the subsequent stages of the design. In an Exploratory design, qualitative information is collected that influences the quantitative methods in later phases of the research design. Conversely, during research that uses an Explanatory research design method, quantitative data is collected that influences the qualitative techniques used in later phases of the research design.

Multiphase Iterative Designs

Multiphase Iterative designs are similar to Exploratory and Explanatory designs in that data is collected in phases that influence later-phase methods used in the design. Typically, a Multiphase Iterative design will use more than two stages in the design; a minimum of three phases is required. The overall process of the design is iterative, meaning researchers traverse between qualitative and quantitative methods as new data is introduced into the study.

Chapter Summary

In this chapter, we learned about the similarities between the Intelligence lifecycle and the Research lifecycle. We also took a closer look at the individual phases of the research lifecycle. In doing so, we found that there is a structured set of stages for which to conduct research and developed a foundational understanding of the fundamentals for research design. We also learned, categorically, that there are several types of research methodologies that influence the sample, data collection, analysis, and results of a study. Lastly, we took a cursory look at some quantitative statistical methods that can be applied when conducting quantitative methods of research. In the next chapter, we will revisit what we had learned previously about the recording of OSINT artifacts. However, this time we will look at how the data in an artifact can be applied to the various types of research methods. Additionally, we will examine some simple examples of applied research methods.

Chapter Acronyms

SPSS: Statistical Package for the Social Sciences

ANOVA: Analysis of Variance

Chapter References

Bar Graphs. (n.d.). Retrieved February 24, 2020, from https://www.mathsisfun.com/data/bar-graphs.html

IBM SPSS Statistics—Overview. (2020, February 21). https://www.ibm.com/products/spss-statistics

Intelligence cycle. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Intelligence_cycle&oldid=937331461

JP 2-0, Joint Intelligence. (2013). 144.

Kent State University Libraries. (2020a, February 10). *LibGuides: SPSS Tutorials: Chi-Square Test of Independence*. https://libguides.library.kent.edu/SPSS/ChiSquare

Kent State University Libraries. (2020b, February 10). *LibGuides: SPSS Tutorials: Crosstabs*. https://libguides.library.kent.edu/SPSS/Crosstabs

Kent State University Libraries. (2020c, February 10). *LibGuides: SPSS Tutorials: Descriptive Stats by Group (Compare Means)*. https://libguides.library.kent.edu/SPSS/CompareMeans

Kent State University Libraries. (2020d, February 10). *LibGuides: SPSS Tutorials: Descriptive Stats for Many Numeric Variables (Descriptives)*. https://libguides.library.kent.edu/SPSS/Descriptives

Kent State University Libraries. (2020e, February 10). *LibGuides: SPSS Tutorials: Descriptive Stats for One Numeric Variable (Explore)*. https://libguides.library.kent.edu/SPSS/Explore

Kent State University Libraries. (2020f, February 10). *LibGuides: SPSS Tutorials: Descriptive Stats for One Numeric Variable (Frequencies)*. https://libguides.library.kent.edu/SPSS/FrequenciesContinuous

Kent State University Libraries. (2020g, February 10). *LibGuides: SPSS Tutorials: Independent Samples t Test*. https://libguides.library.kent.edu/SPSS/IndependentTTest

Kent State University Libraries. (2020h, February 10). *LibGuides: SPSS Tutorials: One Sample t Test*. https://libguides.library.kent.edu/SPSS/OneSampletTest

Kent State University Libraries. (2020i, February 10). *LibGuides: SPSS Tutorials: One-Way ANOVA*. https://libguides.library.kent.edu/SPSS/OneWayANOVA

Kent State University Libraries. (2020j, February 10). *LibGuides: SPSS Tutorials: Paired Samples t Test*. https://libguides.library.kent.edu/SPSS/PairedSamplestTest

Kent State University Libraries. (2020k, February 10). *LibGuides: SPSS Tutorials: Pearson Correlation*. https://libguides.library.kent.edu/SPSS/PearsonCorr

Leedy, P. D., & Ormrod, J. E. (2015). *Practical Research: Planning and Design, Enhanced Pearson eText—Access Card (11th Edition)*. Pearson.

MathIsFun.com. (2017). *Pie Chart*. https://www.mathsisfun.com/data/pie-charts.html

Munro Ph.D., B. H. (2005). *Statistical Methods for Health Care Research (Fifth Edition)* (5th Edition). Lippincott Williams & Wilkins.

Chapter 6: Simulated Application of Research Methods

Chapter 6 Introduction

This chapter is the final chapter for this book. In this chapter, we look at how the previous chapters have built up a level set of knowledge that can be operationalized. Using simulated data, we will see how quantitative methods for research can be leveraged for decision support. Again, these examples are to help analysts frame their cybersecurity research. It is essential to recognize that these examples are not full-blown research proposals or complete studies. The research design architectures are generalized. The efficacy and effectiveness of applying the methods and using the scenarios as foundations for research design will largely depend upon the individual, as well as the access to data and software tailored for research. So, with that, let's dive right in.

Revisiting Frameworks

Now is an excellent time to revisit frameworks before moving further into the bowels of this chapter. In previous chapters, we used frameworks to attach tags to events related to particular OSINT artifacts. Remember, when we discussed quantitative methods, and the term *category* continued to present itself? Well, these tags, created within a hierarchical parent-child relationship, form the basic structure for placing artifact information into categories. We can then develop data mining searches that look for frequencies of tag occurrence, thus making the tags themselves nominal, ordinal, or ratio variable types depending on the data mining criteria. Additionally, by converting the frameworks (MITRE, CWE, CMCF, et al.) into a system taxonomy of using tags, we have also organized our data. Think of the tags as also being a part of a library system used for finding books.

The CMCF

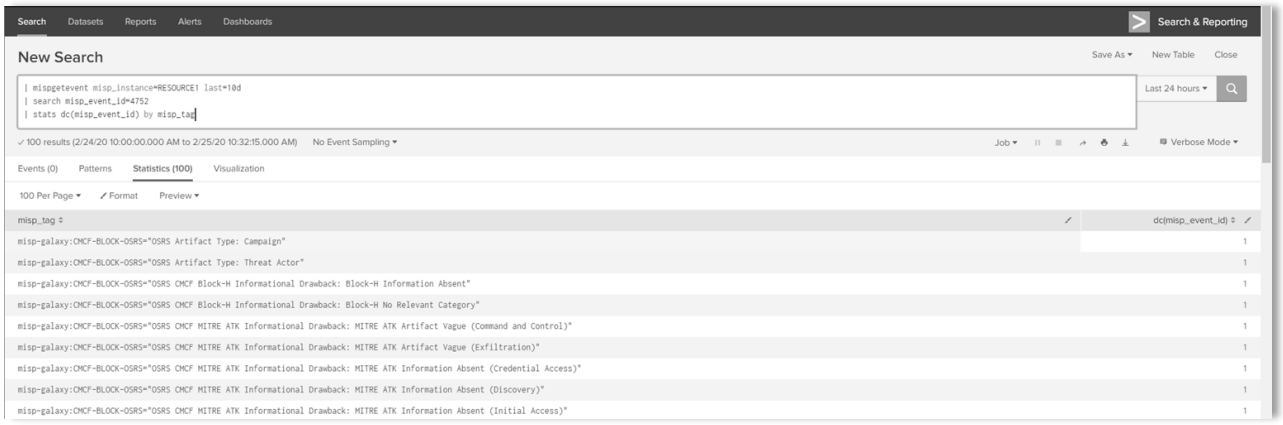
Throughout this book, we discussed several different frameworks. Each one is independent of the other. However, when examining the purpose for each framework, we see that each one by themselves is attempting to describe cyber threat information. MITRE is used for the description of pre-post attack techniques used by a malicious actor. The CWE is used to describe vulnerabilities that can be exploited by the attacker. ENISA is used to describe threats at an operational level that sits between the tactical teams and executive management. Myopically, from the perspective of this book, these frameworks individually do not sufficiently describe threats with a level of completeness that represents a threat in its entirety; and, all at once. The aim of the Comprehensive Modular Cybersecurity Framework (CMCF) is to give those who collect cyber threat information, a mapping that steers them to that sufficient level of completeness. Additionally, the ability to collect information in such a way as to be able to traverse between the individual frameworks easily.

Indicators and Warnings

Because the CMCF tags can be turned into variables that can be analyzed using statistical methods, the CMCF also serves as a tool for developing Indicators and Warnings (IW). Mathematical models that describe a pattern of behavior. For example, MITRE frequencies generated under the CMCF General Artifact category *botnet*. The MITRE frequencies in this example would be considered the indicators. The descriptive statistics that result from the calculation can generate a monitoring process and procedure using descriptive statistical results. If a particular set of behaviors is detected that relates to the MITRE frequencies for the CMCF category *botnet*, an alert, or *warning* can be sent to the appropriate individuals for triage.

Mining for the Data

Data by itself is somewhat meaningless. Data by itself with no ability to access the data is even worse. Luckily where we have stored the data has mechanisms for direct access to the database. MISP, being an open platform with a robust Application Programming Interface (API), innately allows for immediate access to the data stored within its database. Alternate to the API is a direct connection to the database through the use of a Java Database Connectivity (JDBC) connector. The exact engineering for database connectivity is not discussed within the scope of this book. However, we can present an example of what a search of the MISP database might look like when using two tools in conjunction with each other; Splunk and MISP42. The former, Splunk, is a data mining tool that can search through structured and unstructured data using its proprietary search language called the Search Processing Language (SPL). The later, MISP42 is a plugin application for Splunk that allows API access to the MISP database. This combination of software enables the query of the MISP database using Splunk's SPL. The following image is a simple search performed on a single event within MISP that pulls back the tags associated with the event and counts them.



The search that was run looks for a single event with an event ID of 4752 and then counted the occurrence of the event tags by the associated event ID. In this case, the count for each tag is only `1`. However, had this search looked more broadly across all events, the counts of each tag will begin to vary in summation. After we've received the results, we can export the data into a .csv file format to use in another tool or perform additional operations on the data. For example,

we could export this data as a .csv file and then load it into a statistical software platform like SPSS.

Qualitative Scenarios

As discussed in Chapter 5, qualitative studies seek to understand a phenomenon in its natural setting. These types of studies do not place controls on the variables that are measured during the study. Additionally, qualitative studies do not make attempts to reduce the complexity of a phenomenon to break the phenomenon down into simpler parts for interpretation. Finally, qualitative studies do not seek to find a causality from a phenomenon.

Case Study

Scenario

A researcher is looking to understand the interactions between individuals on Darkweb websites that offer the sale of personal information. The communication under observation would is between sellers and sellers/purchasers of personal information.

General Research Design

Case studies require the researcher to observe the phenomenon for a particular duration. For this scenario, the researcher would have to identify one or more Darkweb websites that sell personal information as well as have access to the site. During the period of observation, the researcher would be required to annotate the interactions between the sellers and the seller/purchaser. The annotation tool will most likely be some form of software that allows for the copy and pasting of text into a document as well as take screenshots. Furthermore, during this period of observation, there will be no interaction between the researcher and the observed phenomenon.

Possible Variables

The variables in this study are mainly categorical. Due to the nature of the method, there is no need to categorize the variables into types. However, the following variables can be identified in such a study as well as classified into their particular variable type: Website Title (nominal), Seller Moniker (nominal), Purchaser Moniker (nominal), and Categories of Conversation (discussions on market activity, purchaser reactions, etc.).

Possible Analysis Methods

Analysis of this particular study is not quantitative; therefore, there is no quantitative method applied. Instead, the analysis method would be referential based on researcher annotations, organization of the data, and other relevant pieces of referential evidence that can be used in a results discussion within the final research product.

Expectations of Research

The output of the research should present the findings in a way that conveys the overall complexity of the social dynamics between sellers and sellers/purchasers of personal information on the Darkweb. Furthermore, the output of the research should be detailed enough to lead to a Mixed-Method study that examines these relationships using both qualitative and quantitative methods.

Ethnography

Scenario

A researcher is looking to understand how the language from cybersecurity frameworks is used between similar peer groups. Ethnographic research runs contrary to the goal of a Case Study that seeks to examine the context of the phenomena.

General Research Design

As with the Case Study, analysis for this particular study is not quantitative. Therefore there is no quantitative method applied. By choice, the analysis method would be referential based on researcher annotations, organization of the data, and other relevant pieces of referential evidence that can be used in a discussion section of the final research product. However, the research design should consider the cultural references, habits of usage, colloquial references to frameworks, and how the frameworks are perceived through language.

Possible Variables

Variables for this type of study would initially be categorical and of the variable type nominal. A researcher can assign annotations and observations into these categorical buckets. Categories allow for a further distillation of language into subcategories (e.g., *colloquialisms* or *organizational reference*).

Possible Analysis Methods

Analysis of this particular research would be like the analysis method used by the case study scenario. The analysis process would include these referential categories, along with researcher annotations. These pieces of referential data would be used in a procedure of synthesis portrayed in the final research product.

Expectations of Research

The output of the research should present the findings in a way that conveys the overall depth and the diverse ways for which cybersecurity framework language is used between similar peer groups. Furthermore, the output of the research should be detailed enough to lead to a Mixed-Method study that examines the depth of these using both qualitative and quantitative methods.

Content Analysis

Scenario

A researcher is looking to develop categorical taxonomies that can be applied to a body of knowledge related to internal cybersecurity audits. It has come to the attention of executive leadership that many of the audits report failures within their findings. However, there is no formal organization to the audit results that allows the easy direct focus of resources directed by executive leadership.

General Research Design

Unlike the Case Study or Ethnography scenarios, the research design for Content Analysis does not require the researcher to have any direct access to the phenomenon. Alternatively, the researcher simply needs access to documentation containing the findings of the cybersecurity audits. Once access to the information has been achieved, the researcher can then begin to refine repeated thematical occurrences into their categorical designations.

Possible Variables

The variables for this study would be categorical and of the type nominal. Some possible variables would include but not be limited *department name*, *failure reason,* and *manager name*. These variables could have the potential to transition into ratio variables if the study were to be extended into the quantitative domain of methodologies.

Possible Analysis Methods

One possible analysis method would be a process of classification that breaks the language from the audit findings down into a hierarchical category system. For example, departments might be broken down into their workgroups, so on and so forth. The language would then be observed for thematic occurrences that allow the research to create a system of hierarchical taxonomy for audit findings.

Expectations of Research

The output of the research should present the findings in a way that conveys the observed thematic occurrences as categorical groups. The categorical groups should convey a context of fault causality by the connotation and meaning embedded with the terminology. Furthermore, the output of the research should be detailed enough to lead to a Mixed-Method study that examines the depth of these using both qualitative and quantitative methods.

Quantitative Scenarios

Looking back upon Chapter 5, we see that quantitative studies seek to understand causality through statistical mathematic methods. Different from qualitative studies, researchers who approach their work with quantitative methods in mind know that their research does entirely need to rely on access to people. Quantitative studies have the benefit of being able to have their application applied to living and non-living phenomena. For these qualitative scenarios, we will

not rely on statistical software like SPSS to calculate results. Contrarily, we will simply use a set of publicly available websites and freely available software tailored to perform such calculation. The reason for this is that access to a resource like SPSS by the reader may not be available. Online resources will suffice as we are only looking to understand how a researcher would approach the scenario, determine variables, record data, perform analysis, and then interpret the results. All websites used for these scenarios can be found within the *references* section of this chapter. Last, the data used is simply data generated for demonstration purposes only and does not reflect the possible results from actual OSINT data recorded from publicly available sources.

Descriptive Statistics and Central Tendency

Scenario

A researcher is looking to determine the central tendencies of OSINT botnet artifacts classified using the CMCF mapping. More specifically, the researcher is looking to identify central tendencies for botnets as they relate to the MITRE framework.

General Research Design

In this scenario, we are assuming that the researcher has access to both the CMCF and a way to assign a MITRE framework classification to botnet related OSINT artifacts. We will continue to assume that the researcher is using MISP to record artifacts. Over time the researcher will collect a sufficient number of artifacts (more than ten), record, and attach the appropriate MITRE framework tags. These tags will then be counted as observations of occurrence; or frequencies.

Possible Variables

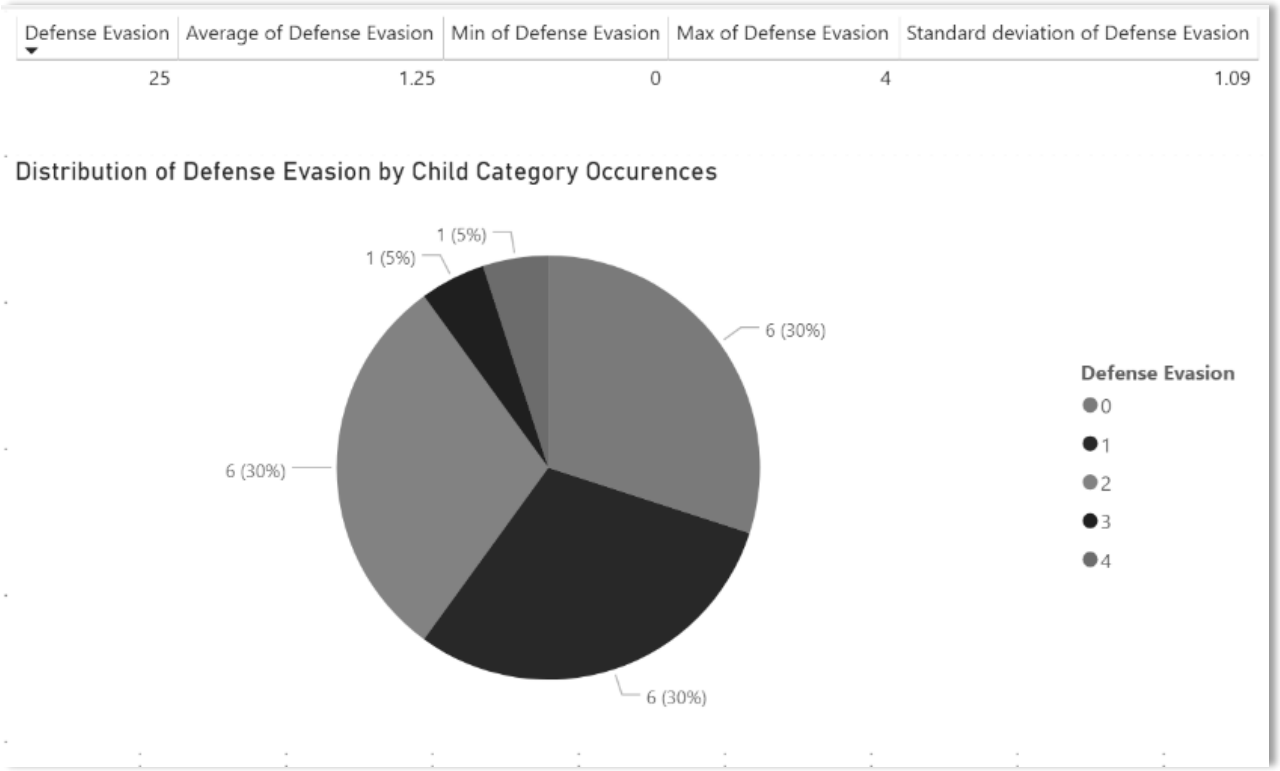
Descriptive statistics and central tendency do not require particular variable types like nominal, ordinal, or ratio. Instead, the variables simply need to have the ability to be summarized numerically. In this scenario, we have two parent hierarchical categories *Artifact Type* and *MITRE ATK*. There is only one child category being used beneath the parent category Artifact Type, and that child category is *Botnet*. Under MITRE, we have several other child categories that can be examined. *Initial Access, Defense Evasion,* and *Collection* are just a few of the possible child categories under MITRE ATK. For this scenario, we will focus on developing descriptive statistics and central tendencies for the MITRE child category Defense Evasion.

Possible Analysis Methods

In this scenario, we are looking at the occurrences of sub-child category tags. Discovery of occurrence means we are looking at how many times a set of tags occurred over a collection of artifacts, and what those summations were. These summations make up the frequencies within the individual event, and together make a series of frequencies that can be used to find the central tendencies.

Expectations of Research

This research expects to develop a statistically developed mathematical profile of central tendencies for the observed phenomenon botnet. The output from this research should show at a minimum the mean (or average), the total summation, the minimum value of the data set, the maximum value of the data set, and the standard deviation. The output should also be able to portray these results visually using a chart (pie, bar, line, et al.).



Using Microsoft's Power BI, which is available for download at no cost, we can see that the mean for Defense Evasion is 1.25, the minimum value is 0, the maximum is 4, and has a standard deviation of 1.09. We can also see from the pie chart that Defense Evasion frequency values of 0 and 1 make up 60% of the total Defense Evasion values; 30% each. The values can be interpreted as over half of all botnet artifacts sampled; of those, 60% displayed little-to-no Defense Evasion MITRE ATK techniques. These values can further be understood from an IW viewpoint that portrays botnets as malicious attacks that do not make attempts at evading detection often. Therefore, as a decision support recommendation, a cyber intelligence analyst may recommend it better to monitor for other types of botnet attack techniques with higher levels of frequency since statistics show they often do not attempt to evade detection.

Person Correlation

Scenario

A researcher is looking to understand the direction of a linear relationship between the number of malicious websites recorded per day and the number of true positive cybersecurity events that required a triage response, as well as the strength of this relationship.

General Research Design

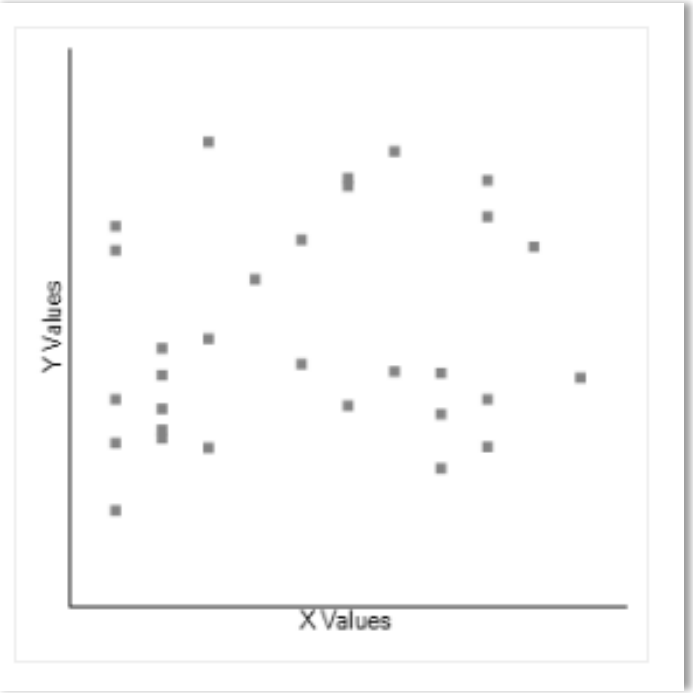
For this research design, the researcher would need access to systems that generate data in a way that classifies a website as malicious or not. Additionally, the researcher would need access to triage data for the organization. We will assume for this study that the researcher has access to both sets of data and has been collecting this data for thirty days.

Possible Variables

The independent variable in this study would is the malicious websites (variable type ratio) since we are examining the frequency of this category, and a frequency of '0' would mean that no website was visited that day; hence an absence of value. The dependent variable for this study is the number of true positive cybersecurity events triaged per day. Again, the dependent variable is a ratio variable in the same manner as the independent variable.

Possible Analysis Methods

For this study, we are explicitly using the quantitative method, Person Correlation. We will not be examining the central tendencies for each variable. We will also use the publicly available resource *Social Science Statistics* and the available Person Correlation calculator provided by the website.



Expectations of Research

This research expects to determine the Person Correlation Coefficient. The correlation coefficient allows us to understand the correlation and strength of the relationship between the two variables. From our simulation data, we see that the *R*-value (correlation coefficient) is 0.1733. This value is relatively close to a value of '0' and can be interpreted as having a weak relationship between the variables. Remembering that a correlation coefficient ranges between '-1', '0', and '1', with a perfect zero interpreted as absolutely no correlation or

relationship. We can also represent the results with a scatter plot.

Though the *R*-value is too low to say that there is enough of a relationship to use the variables as causality predictors, it does show that there is a relationship between the number of malicious sites visited per day and the number of true positive triage events. As a matter of decision support, this might be enough information for leadership to conduct efforts at reducing the number of malicious sites visited per day by the user population.

Paired Sample t Test

Scenario

For this scenario, we look back to a previous example for pretest-posttest effects of rule tuning and the number of false positives generated by a cybersecurity monitoring tool over thirty days. However, this time we will expand on the example and show how simulated data is represented in a Paired Sample t Test.

General Research Design

This research design considers a single system that can classify monitored data as either a true positive cybersecurity event or as a false positive cybercity event. The researcher will simply collect the pretest data before the rule tune. Then, once the pretest data has been collected, the researcher will implement rule tuning and collect another set of sample data. Additionally, this quantitative method is used in hypothesis testing. Therefore, before conducting the test, we must set up both the null (H_1) and alternative (H_2) hypothesis.

H_1: there is no change to the frequency in reported false positive events as a result of rule tuning

H_2: there is a change to the frequency in reported false positive events as a result of rule tuning

Possible Variables

In this study, the dependent variable is the number of false positives reported by the system, a variable of type ratio. The independent variable is the variable type nominal expressed by the categories pretest and posttest.

Possible Analysis Methods

For this study, we are explicitly using the quantitative method Paired Sample t Test. Since this type of statistical method provides a p-value as part of the results, and we will set our *alpha* value at .05. This alpha value will be used to accept or reject the null hypothesis. We will not be examining the central tendencies for each variable. We will also use the publicly available resource *Statistics Kingdom* and the available Paired Sample t Test calculator provided by the website.

Expectations of Research

This research expects to determine the p-value and either accept or reject the null hypothesis. Using simulation data, the p-value from the test is 0.317479. Therefore, for this particular scenario, based on the results from the analysis, the researcher would accept the null hypothesis (H_1). However, this does not mean that rule tuning did not affect the number of false positives. Simply, the effort did not have enough of an effect to show that rule tuning had a statistically significant impact on the number of false positives.

Experimental Scenarios

One-Shot Experimental Case Study

Scenario

A researcher is looking to see if televising cybersecurity awareness content regarding malicious emails on organizational video systems results in a significant increase in the number of potentially malicious emails reported by corporate employees to the cybersecurity operations team that triages these submissions.

General Research Design

Since this type of experimental method does not attempt to control for confounding variables, the researcher simply needs a pretest-posttest mean that can be analyzed for statistical significance. The researcher will also need to create the cybersecurity awareness content, have access to the dissemination mechanism, and access to the number of reported potentially malicious emails. This study can be conducted over a thirty-day time span that is broken into two halves for sampling purposes. A Paired Sample t Test will be used to compare the means from the pretest-posttest data. Because a Paired Sample t Test is being used, we must also include the null and alternate hypothesis statements.

H_1: there is no change to the frequency in reported malicious spam to the cybersecurity operations team

H_2: there is a change to the frequency in reported malicious spam to the cybersecurity operations team

Possible Variables

In this scenario, the dependent variable is the number of potentially malicious spam emails reported to the cybersecurity operations team, a variable of type ratio. The independent variable is the variable type nominal expressed by the categories pretest and posttest.

Possible Analysis Methods

For this scenario, we are explicitly using the quantitative method Paired Sample t Test incorporated into the experimental research design. Since this type of statistical method provides

a p-value as a part of the results, we will set our *alpha* value at .05. The p-value will be used to accept or reject the null hypothesis. We will not be examining the central tendencies for each variable. We will also use the publicly available resource *Statistics Kingdom* and the available Paired Sample t Test calculator provided by the website.

Expectations of Research

The output expectations of this research will be the p-value that either allows the researcher to accept or reject the null hypothesis. Using simulation data, the p-value from the test is 0.398506. Therefore, for this particular scenario, based on the results from the analysis, the researcher would accept the null hypothesis (H_1). However, this does not mean that the cybersecurity awareness content did not affect the sample, nor does this type of research design consider any confounding variables. As an example of a confounding variable, if the researcher designed the cybersecurity awareness content without informing the sample of how to contact the cybersecurity operations team with possibly malicious emails, the lack of contact information would affect the total number of reports to cybersecurity operations.

Simple Time-Series Design

Scenario

For this scenario, we will continue to build upon the introduction of treatment seen in the previous scenario regarding the reporting of potentially malicious spam. However, there is a change in design architecture. We will be conducting a Simple Time-Series design and comparing the means to determine if the treatment posed statistical significance or not. The most exciting part of this scenario is the fact that it attempts to see what the effect the confounding variable of time has on the sample.

General Research Design

The researcher will use the same cybersecurity awareness content from the previous scenario. The reused cybersecurity awareness content will act as a type of pseudo control that is carried over from the One-Shot Experimental Case study. We know that this content has not changed and has remained constant. The study will be conducted over four consecutive sampling periods of fifteen days each (N = 15). As a change to the design that differs from the One-Shot Experimental Case study, there will be a baseline created in the first interval. Here, the treatment (cybersecurity awareness content) will be introduced to the sample. The second, third, and four intervals will be compared against this baseline.

Possible Variables

The dependent variable is the number of potentially malicious spam emails reported to the cybersecurity operations team, a variable of type ratio. The independent variable is the variable type nominal expressed by the categories pretest and posttests (e.g., second, third, and fourth sampling).

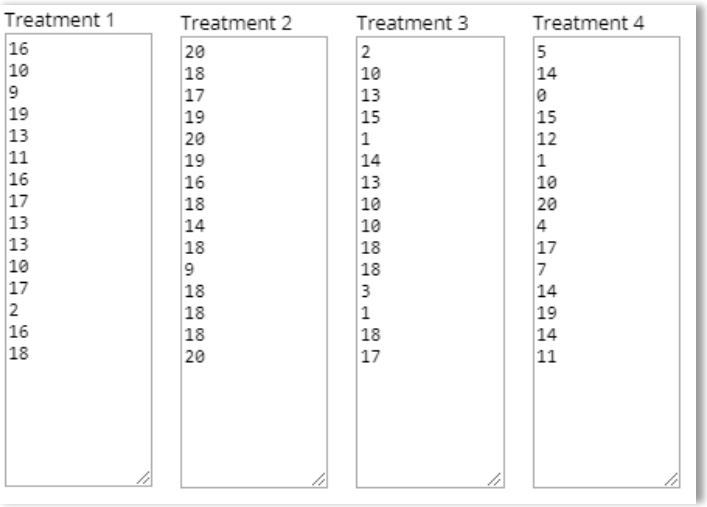
Possible Analysis Methods

We will use a One-Way ANOVA analysis to compare more than two means to determine statistical significance. Because a One-Way ANOVA reports a p-value, we must set the null and alternative hypothesis, as well as the alpha value of .05 before testing. We will be using the publicly available online resource for calculating a One-Way ANOVA found at the website *Social Science Statistics*. Once more, we will be using simulated data for this test.

H_1: there is no change to the frequency in reported malicious spam to the cybersecurity operations team

H_2: there is a change to the frequency in reported malicious spam to the cybersecurity operations team

Expectations of Research



From our test, we see that the *f*-ratio was 5.4, and the p-value was .002. The p-value is below the .05 preset alpha value. Therefore, the analysis shows statistical significance, and the researcher will reject the null hypothesis and accept the alternate hypothesis. As an interpretation of the results, we can say that the introduction of the treatment did have an effect on the sample and was statistically significant.

Summary of Data

	Treatments					
	1	2	3	4	5	Total
N	15	15	15	15		60
$\sum X$	200	262	163	163		788
Mean	13.3333	17.4667	10.8667	10.8667		13.133
$\sum X^2$	2944	4688	2335	2319		12286
Std.Dev.	4.4508	2.8251	6.3456	6.2549		5.7297

Historical Scenarios

Primary Source Synthesis

Scenario

A researcher is looking to synthesize the phenomenon data breach into a concise history of events for a year.

General Research Design

The researcher will use publicly available OSINT artifacts that convey information regarding data breach events. Possible secondary sources of previous historical studies of the data breach phenomenon may be included in the body of data if the researcher deems them a necessary part of the research effort.

Possible Variables

Because this is a historical method, and the goal of the research is to synthesize the information into a concise body of knowledge, there are no variables used within this study.

Possible Analysis Methods

As a general method of organization, the researcher may categorize the information in a way that reflects the organization of presentation in the final research product.

Expectations of Research

The research is expected to create a synthesized body of phenomenological knowledge regarding *a specific period of time*.

Mixed-Method Scenario

A Study Data Breach and Geolocation

Scenario

A researcher is looking to gain further insight into the phenomenon *Data Breach* events and the phenomenological relationship to geolocation, meaning, is there a relationship between data breaches and where they occur.

General Research Design

The specific Mixed-Method for this study is and Embedded Mixed-Method design. The first phase, a qualitative phase, will encompass the collection of data and analyzed using qualitative methods. The data captured during the qualitative phase will be used to enrich the quantitative portion of the study. The two primary methods used in this study are the qualitative method

Content Analysis, and minimal qualitative methods of *Descriptive Statistics*; as well as, the application of Paired Sample t Tests and One-Way ANOVA where applicable. The researcher will use publicly available sources to gather information regarding data breach events. These events will be recorded into MISP and tagged using the CMCF mapping. Every artifact recorded will be tagged with a minimum of artifact type, cause of the breach, location of breach by U.S. state/Non-U.S. state, and all artifacts will have a data breach severity score assigned. Frequencies are developed through tag occurrence within the database. Splunk will be used to mine the data within the MISP database. SPSS will be used for statistical analysis.

Possible Variables

Several categorical variables can either be examined as a type nominal or type ratio variable. For example, the cause of the breach and location of the breach (U.S. state/Non-U.S. state) can be treated this way.

Possible Analysis Methods

Content Analysis will consider the textual information regarding the data breach events to synthesize the phenomenon into an organized body of knowledge. Descriptive statistics, like frequency tables, can be developed using the frequency of tag occurrences. Quantitatively the Chi-Square Test of Independence, Person Correlation, and One-Way ANOVA can also be applied to the type ratio variables created from the categories and frequencies from the attached tags to the event.

Expectations of Research

The research expects to collect enough openly available data regarding the phenomenon *Data Breach* that allows the researcher to create a final product that examines the phenomenon from both qualitative and quantitative perspectives. The final research product will include discussions regarding the results from both the qualitative and quantitative methods.

Chapter Summary

In this chapter, we laid out some example scenarios using the methods described in Chapter 5. In truth, these example scenarios only touched the surface of applied research for cybersecurity. The goal of the chapter was to give enough examples with a level of detail that provides an analyst with a glimpse of practical method application. However, the scenarios did not take the reader into a depth of detail that portrays a more comprehensive approach to research. In this chapter, we did not discuss the importance of a Literature Review, nor did we discuss the development of a Research Proposal. Both the Literature Review and Research Proposal add additional rigor that enhances the research design with extended validity and purpose. Hopefully, in subsequent editions of this book, we will take a deeper dive that incorporates both into the body of knowledge recorded here.

Chapter Acronyms

API: Application Programming Interface

CMCF: Comprehensive Modular Cybersecurity Framework

CSV: Comma-separated Values

ID: Identification

IW: Indicators and Warning

JDBC: Java Database Connectivity

OSINT: Open Source Intelligence

SPL: Search Processing Language

SPSS: Statistical Package for the Social Sciences

https://www.linkedin.com/company/threathunting
https://www.twitter.com/threathunting_

Chapter References

Application programming interface. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Application_programming_interface&oldid=941698972

Common Weakness Enumeration. (2019). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Common_Weakness_Enumeration&oldid=918897150

Data Visualization | Microsoft Power BI. (n.d.). Retrieved February 25, 2020, from https://powerbi.microsoft.com/en-us/

ENISA. (n.d.). Retrieved February 4, 2020, from https://www.enisa.europa.eu/

Grabo, C. M. (2004). *Anticipating Surprise: Analysis for Strategic Warning*. University of America Press, Inc.

IBM SPSS Statistics—Overview. (2020, February 21). https://www.ibm.com/products/spss-statistics

Java Database Connectivity. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Java_Database_Connectivity&oldid=938614427

Leedy, P. D., & Ormrod, J. E. (2015). *Practical Research: Planning and Design, Enhanced Pearson eText—Access Card (11th Edition)*. Pearson.

MISP - Open Source Threat Intelligence Platform & Open Standards For Threat Information Sharing (formely known as Malware Information Sharing Platform). (n.d.). Retrieved February 4, 2020, from https://www.misp-project.org/

MITRE ATT&CK™. (n.d.). Retrieved February 18, 2020, from https://attack.mitre.org/

Munro Ph.D., B. H. (2005). *Statistical Methods for Health Care Research (Fifth Edition)* (5th Edition). Lippincott Williams & Wilkins.

Open-source intelligence. (2020). In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Open-source_intelligence&oldid=937269341

open-source-rs. (2020). *Open-source-rs/Comprehensive-Modular-Cybersecurity-Framework-CMCF*. https://github.com/open-source-rs/Comprehensive-Modular-Cybersecurity-Framework-CMCF (Original work published 2019)

Paired T-Test calculator. (n.d.-a). Retrieved February 25, 2020, from http://www.statskingdom.com/160MeanT2pair.html

Paired T-Test calculator. (n.d.-b). Retrieved February 25, 2020, from http://www.statskingdom.com/160MeanT2pair.html

Pearson Correlation Coefficient Calculator. (n.d.). Retrieved February 25, 2020, from https://www.socscistatistics.com/tests/pearson/

Search Processing Language | Resources | Splunk. (n.d.). Retrieved February 25, 2020, from https://www.splunk.com/en_us/resources/search-processing-language.html

Séguy, R. (2020). *Remg427/misp42splunk* [Python]. https://github.com/remg427/misp42splunk (Original work published 2017)

SIEM, AIOps, Application Management, Log Management, Machine Learning, and Compliance. (n.d.). Splunk. Retrieved February 25, 2020, from https://www.splunk.com/en_us/homepage.html

GNU Free Documentation License
Version 1.3, 3 November 2008

0. PREAMBLE
The purpose of this License is to make a manual, textbook, or other functional and useful document "free" in the sense of freedom: to assure everyone the effective freedom to copy and redistribute it, with or without modifying it, either commercially or noncommercially. Secondarily, this License preserves for the author and publisher a way to get credit for their work, while not being considered responsible for modifications made by others.

This License is a kind of "copyleft", which means that derivative works of the document must themselves be free in the same sense. It complements the GNU General Public License, which is a copyleft license designed for free software.

We have designed this License in order to use it for manuals for free software, because free software needs free documentation: a free program should come with manuals providing the same freedoms that the software does. But this License is not limited to software manuals; it can be used for any textual work, regardless of subject matter or whether it is published as a printed book. We recommend this License principally for works whose purpose is instruction or reference.

1. APPLICABILITY AND DEFINITIONS
This License applies to any manual or other work, in any medium, that contains a notice placed by the copyright holder saying it can be distributed under the terms of this License. Such a notice grants a world-wide, royalty-free license, unlimited in duration, to use that work under the conditions stated herein. The "Document", below, refers to any such manual or work. Any member of the public is a licensee, and is addressed as "you". You accept the license if you copy, modify or distribute the work in a way requiring permission under copyright law.

A "Modified Version" of the Document means any work containing the Document or a portion of it, either copied verbatim, or with modifications and/or translated into another language.

A "Secondary Section" is a named appendix or a front-matter section of the Document that deals exclusively with the relationship of the publishers or authors of the Document to the Document's overall subject (or to related matters) and contains nothing that could fall directly within that overall subject. (Thus, if the Document is in part a textbook of mathematics, a Secondary Section may not explain any mathematics.) The relationship could be a matter of historical connection with the subject or with related matters, or of legal, commercial, philosophical, ethical or political position regarding them.

The "Invariant Sections" are certain Secondary Sections whose titles are designated, as being those of Invariant Sections, in the notice that says that the Document is released under this License. If a section does not fit the above definition of Secondary then it is not allowed to be designated as Invariant. The Document may contain zero Invariant Sections. If the Document does not identify any Invariant Sections then there are none.

The "Cover Texts" are certain short passages of text that are listed, as Front-Cover Texts or Back-Cover Texts, in the notice that says that the Document is released under this License. A Front-Cover Text may be at most 5 words, and a Back-Cover Text may be at most 25 words.

A "Transparent" copy of the Document means a machine-readable copy, represented in a format whose specification is available to the general public, that is suitable for revising the document straightforwardly with generic text editors or (for images composed of pixels) generic paint programs or (for drawings) some widely available drawing editor, and that is suitable for input to text formatters or for automatic translation to a variety of formats suitable for input to text formatters. A copy made in an otherwise Transparent file format whose markup, or absence of markup, has been arranged to thwart or discourage subsequent modification by readers is not Transparent. An image format is not Transparent if used for any substantial amount of text. A copy that is not "Transparent" is called "Opaque".

Examples of suitable formats for Transparent copies include plain ASCII without markup, Texinfo input format, LaTeX input format, SGML or XML using a publicly available DTD, and standard-conforming simple HTML, PostScript or PDF designed for human modification. Examples of transparent image formats include PNG, XCF and JPG. Opaque formats include proprietary formats that can be read and edited only by proprietary word processors, SGML or XML for which the DTD and/or processing tools are not generally available, and the machine-generated HTML, PostScript or PDF produced by some word processors for output purposes only.

The "Title Page" means, for a printed book, the title page itself, plus such following pages as are needed to hold, legibly, the material this License requires to appear in the title page. For works in formats which do not have any title page as such, "Title Page" means the text near the most prominent appearance of the work's title, preceding the beginning of the body of the text.

The "publisher" means any person or entity that distributes copies of the Document to the public.

A section "Entitled XYZ" means a named subunit of the Document whose title either is precisely XYZ or contains XYZ in parentheses following text that translates XYZ in another language. (Here XYZ stands for a specific section name mentioned below, such as "Acknowledgements", "Dedications", "Endorsements", or "History".) To "Preserve the Title" of such a section when you modify the Document means that it remains a section "Entitled XYZ" according to this definition.

The Document may include Warranty Disclaimers next to the notice which states that this License applies to the Document. These Warranty Disclaimers are considered to be included by reference in this License, but only as regards disclaiming warranties: any other implication that these Warranty Disclaimers may have is void and has no effect on the meaning of this License.

2. VERBATIM COPYING
You may copy and distribute the Document in any medium, either commercially or noncommercially, provided that this License, the copyright notices, and the license notice saying this License applies to the Document are reproduced in all copies, and that you add no other conditions whatsoever to those of this License. You may not use technical measures to obstruct or control the reading or further copying of the copies you make or distribute. However, you may accept compensation in exchange for copies. If you distribute a large enough number of copies you must also follow the conditions in section 3.

You may also lend copies, under the same conditions stated above, and you may publicly display copies.

3. COPYING IN QUANTITY
If you publish printed copies (or copies in media that commonly have printed covers) of the Document, numbering more than 100, and the Document's license notice requires Cover Texts, you must enclose the copies in covers that carry, clearly and legibly, all these Cover Texts: Front-Cover Texts on the front cover, and Back-Cover Texts on the back cover. Both covers must also clearly and legibly identify you as the publisher of these copies. The front cover must present the full title with all words of the title equally prominent and visible. You may add other material on the covers in addition. Copying with changes limited to the covers, as long as they preserve the title of the Document and satisfy these conditions, can be treated as verbatim copying in other respects.

If the required texts for either cover are too voluminous to fit legibly, you should put the first ones listed (as many as fit reasonably) on the actual cover, and continue the rest onto adjacent pages.

If you publish or distribute Opaque copies of the Document numbering more than 100, you must either include a machine-readable Transparent copy along with each Opaque copy, or state in or with each Opaque copy a computer-network location from which the general network-using public has access to download using public-standard network protocols a complete Transparent copy of the Document, free of added material. If you use the latter option, you must take reasonably prudent steps, when you begin distribution of Opaque copies in quantity, to ensure that this Transparent copy will remain thus accessible at the stated location until at least one year after the last time you distribute an Opaque copy (directly or through your agents or retailers) of that edition to the public.

It is requested, but not required, that you contact the authors of the Document well before redistributing any large number of copies, to give them a chance to provide you with an updated version of the Document.

4. MODIFICATIONS
You may copy and distribute a Modified Version of the Document under the conditions of sections 2 and 3 above, provided that you release the Modified Version under precisely this License, with the Modified Version filling the role of the Document, thus licensing distribution and modification of the Modified Version to whoever possesses a copy of it. In addition, you must do these things in the Modified Version:

A. Use in the Title Page (and on the covers, if any) a title distinct from that of the Document, and from those of previous versions (which should, if there were any, be listed in the History section of the Document). You may use the same title as a previous version if the original publisher of that version gives permission.
B. List on the Title Page, as authors, one or more persons or entities responsible for authorship of the modifications in the Modified Version, together with at least five of the principal authors of the Document (all of its principal authors, if it has fewer than five), unless they release you from this requirement.
C. State on the Title page the name of the publisher of the Modified Version, as the publisher.
D. Preserve all the copyright notices of the Document.
E. Add an appropriate copyright notice for your modifications adjacent to the other copyright notices.
F. Include, immediately after the copyright notices, a license notice giving the public permission to use the Modified Version under the terms of this License, in the form shown in the Addendum below.
G. Preserve in that license notice the full lists of Invariant Sections and required Cover Texts given in the Document's license notice.
H. Include an unaltered copy of this License.
I. Preserve the section Entitled "History", Preserve its Title, and add to it an item stating at least the title, year, new authors, and publisher of the Modified Version as given on the Title Page. If there is no section Entitled "History" in the Document, create one stating the title, year, authors, and publisher of the Document as given on its Title Page, then add an item describing the Modified Version as stated in the previous sentence.
J. Preserve the network location, if any, given in the Document for public access to a Transparent copy of the Document, and likewise the network locations given in the Document for previous versions it was based on. These may be placed in the "History" section. You may omit a network location for a work that was published at least four years before the Document itself, or if the original publisher of the version it refers to gives permission.
K. For any section Entitled "Acknowledgements" or "Dedications", Preserve the Title of the section, and preserve in the section all the substance and tone of each of the contributor acknowledgements and/or dedications given therein.
L. Preserve all the Invariant Sections of the Document, unaltered in their text and in their titles. Section numbers or the equivalent are not considered part of the section titles.
M. Delete any section Entitled "Endorsements". Such a section may not be included in the Modified Version.
N. Do not retitle any existing section to be Entitled "Endorsements" or to conflict in title with any Invariant Section.
O. Preserve any Warranty Disclaimers.
If the Modified Version includes new front-matter sections or appendices that qualify as Secondary Sections and contain no material copied from the Document, you may at your option designate some or all of these sections as invariant. To do this, add their titles to the list of Invariant Sections in the Modified Version's license notice. These titles must be distinct from any other section titles.

You may add a section Entitled "Endorsements", provided it contains nothing but endorsements of your Modified Version by various parties—for example, statements of peer review or that the text has been approved by an organization as the authoritative definition of a standard.

You may add a passage of up to five words as a Front-Cover Text, and a passage of up to 25 words as a Back-Cover Text, to the end of the list of Cover Texts in the Modified Version. Only one passage of Front-Cover Text and one of Back-Cover Text may be added by (or through arrangements made by) any one entity. If the Document already includes a cover text for the same cover, previously added by you or by arrangement made by the same entity you are acting on behalf of, you may not add another; but you may replace the old one, on explicit permission from the previous publisher that added the old one.

The author(s) and publisher(s) of the Document do not by this License give permission to use their names for publicity for or to assert or imply endorsement of any Modified Version.

5. COMBINING DOCUMENTS
You may combine the Document with other documents released under this License, under the terms defined in section 4 above for modified versions, provided that you include in the combination all of the Invariant Sections of all of the original documents, unmodified, and list them all as Invariant Sections of your combined work in its license notice, and that you preserve all their Warranty Disclaimers.

The combined work need only contain one copy of this License, and multiple identical Invariant Sections may be replaced with a single copy. If there are multiple Invariant Sections with the same name but different contents, make the title of each such section unique by adding at the end of it, in parentheses, the name of the original author or publisher of that section if known, or else a unique number. Make the same adjustment to the section titles in the list of Invariant Sections in the license notice of the combined work.

In the combination, you must combine any sections Entitled "History" in the various original documents, forming one section Entitled "History"; likewise combine any sections Entitled "Acknowledgements", and any sections Entitled "Dedications". You must delete all sections Entitled "Endorsements".

6. COLLECTIONS OF DOCUMENTS
You may make a collection consisting of the Document and other documents released under this License, and replace the individual copies of this License in the various documents with a single copy that is included in the collection, provided that you follow the rules of this License for verbatim copying of each of the documents in all other respects.

You may extract a single document from such a collection, and distribute it individually under this License, provided you insert a copy of this License into the extracted document, and follow this License in all other respects regarding verbatim copying of that document.

7. AGGREGATION WITH INDEPENDENT WORKS
A compilation of the Document or its derivatives with other separate and independent documents or works, in or on a volume of a storage or distribution medium, is called an "aggregate" if the copyright resulting from the compilation is not used to limit the legal rights of the compilation's users beyond what the individual works permit. When the Document is included in an aggregate, this License does not apply to the other works in the aggregate which are not themselves derivative works of the Document.

If the Cover Text requirement of section 3 is applicable to these copies of the Document, then if the Document is less than one half of the entire aggregate, the Document's Cover Texts may be placed on covers that bracket the Document within the aggregate, or the electronic equivalent of covers if the Document is in electronic form. Otherwise they must appear on printed covers that bracket the whole aggregate.

8. TRANSLATION
Translation is considered a kind of modification, so you may distribute translations of the Document under the terms of section 4. Replacing Invariant Sections with translations requires special permission from their copyright holders, but you may include translations of some or all Invariant Sections in addition to the original versions of these Invariant Sections. You may include a translation of this License, and all the license notices in the Document, and any Warranty Disclaimers, provided that you also include the original English version of this License and the original versions of those notices and disclaimers. In case of a disagreement between the translation and the original version of this License or a notice or disclaimer, the original version will prevail.

If a section in the Document is Entitled "Acknowledgements", "Dedications", or "History", the requirement (section 4) to Preserve its Title (section 1) will typically require changing the actual title.

9. TERMINATION
You may not copy, modify, sublicense, or distribute the Document except as expressly provided under this License. Any attempt otherwise to copy, modify, sublicense, or distribute it is void, and will automatically terminate your rights under this License.

However, if you cease all violation of this License, then your license from a particular copyright holder is reinstated (a) provisionally, unless and until the copyright holder explicitly and finally terminates your license, and (b) permanently, if the copyright holder fails to notify you of the violation by some reasonable means prior to 60 days after the cessation.

Moreover, your license from a particular copyright holder is reinstated permanently if the copyright holder notifies you of the violation by some reasonable means, this is the first time you have received notice of violation of this License (for any work) from that copyright holder, and you cure the violation prior to 30 days after your receipt of the notice.

Termination of your rights under this section does not terminate the licenses of parties who have received copies or rights from you under this License. If your rights have been terminated and not permanently reinstated, receipt of a copy of some or all of the same material does not give you any rights to use it.

10. FUTURE REVISIONS OF THIS LICENSE
The Free Software Foundation may publish new, revised versions of the GNU Free Documentation License from time to time. Such new versions will be similar in spirit to the present version, but may differ in detail to address new problems or concerns. See https://www.gnu.org/licenses/.

Each version of the License is given a distinguishing version number. If the Document specifies that a particular numbered version of this License "or any later version" applies to it, you have the option of following the terms and conditions either of that specified version or of any later version that has been published (not as a draft) by the Free Software Foundation. If the Document does

not specify a version number of this License, you may choose any version ever published (not as a draft) by the Free Software Foundation. If the Document specifies that a proxy can decide which future versions of this License can be used, that proxy's public statement of acceptance of a version permanently authorizes you to choose that version for the Document.

11. RELICENSING
"Massive Multiauthor Collaboration Site" (or "MMC Site") means any World Wide Web server that publishes copyrightable works and also provides prominent facilities for anybody to edit those works. A public wiki that anybody can edit is an example of such a server. A "Massive Multiauthor Collaboration" (or "MMC") contained in the site means any set of copyrightable works thus published on the MMC site.

"CC-BY-SA" means the Creative Commons Attribution-Share Alike 3.0 license published by Creative Commons Corporation, a not-for-profit corporation with a principal place of business in San Francisco, California, as well as future copyleft versions of that license published by that same organization.

"Incorporate" means to publish or republish a Document, in whole or in part, as part of another Document.

An MMC is "eligible for relicensing" if it is licensed under this License, and if all works that were first published under this License somewhere other than this MMC, and subsequently incorporated in whole or in part into the MMC, (1) had no cover texts or invariant sections, and (2) were thus incorporated prior to November 1, 2008.

The operator of an MMC Site may republish an MMC contained in the site under CC-BY-SA on the same site at any time before August 1, 2009, provided the MMC is eligible for relicensing.