Project Deliverable

D8.2
Data Management Plan

Project Number	700692
Project Title	DiSIEM – Diversity-enhancements for SIEMs
Programme	H2020-DS-04-2015

Deliverable type	Report
Dissemination level	PU
Submission date	28.02.2017

Responsible partner	EDP
Editor	Pedro Dias Rodrigues
Revision	0.11

The DiSIEM project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 700692.

Editor
Pedro Dias Rodrigues, EDP

Contributors
Ana Respício, FFCUL
João Alves, FFCUL
Luís Miguel Ferreira, FFCUL
Alysson Bessani, FFCUL
Pedro Dias Rodrigues, EDP
Gonçalo Santos Martins, EDP
Ivo Rosa, EDP
Susana González Zarzosa, Atos
Michael Kamp, Fraunhofer
Zayani Dabbabi, AMADEUS
Ilir Gashi, City

Executive Summary

This document constitutes the Data Management Plan (DMP) of the DiSIEM project, explaining how the project plans to manage datasets. It records and foresees the activities of all DiSIEM partners related to the production and use of datasets (for experimentation, measurement, demonstration or validation purposes).

The document has been compiled as a summary of a questionnaire-based survey distributed to all the partners of the DiSIEM consortium.

There are two key types of data being produced in the scope of the DiSIEM project: raw data originated in the SIEM systems operated by industry partners (Tier 1) and processed data generated by the remaining partners (Tier 2).

The most prevalent data format is Comma-Separated Values (CSV), a textual description of data that is common and widely used in the big data community. Several partners intend to share their datasets publicly, making them available for further research outside the project.

Since it is very early in the project, little is known in terms of sharing, volume and archiving. The project is aware of these aspects and will tackle them by updating the present document during the development of the specifications of the experimentations. Therefore, information in this document is subject to change.

Table of Contents

List of Tables

Revision History

Version	Date	Author	Notes
0.1	29.12.2016	Pedro Dias Rodrigues (EDP)	First contribution.
0.2	25.01.2017	Pedro Dias Rodrigues (EDP)	Review of document contents. Integration of comments by other partners.
0.3	07.02.2017	Pedro Dias Rodrigues (EDP)	Added input for EDP dataset
0.4	09.02.2017	Michael Kamp (Fraunhofer)	Added input for Fraunhofer dataset
0.5	14.02.2017	Susana González Zarzosa (Atos)	Added input for Atos dataset
0.6	16.02.2017	Ilir Gashi (City)	Added input for City dataset
0.7	19.02.2017	Alysson Neves Bessani (FFCUL)	Added input on FFCUL dataset
0.8	20.02.2017	Zayani Dabbabi (AMADEUS)	Added input for AMADEUS dataset
0.9	24.02.2017	Pedro Dias Rodrigues (EDP)	Document review
0.10	27.02.2017	Alysson Neves Bessani (FFCUL)	Document review
0.11	28.02.2017	Pedro Dias Rodrigues (EDP)	Document review
1.0	02.03.2017	Pedro Dias Rodrigues (EDP)	First version after final review

1 Introduction

The Commission is running a flexible pilot under Horizon 2020 called the Open Research Data (ORD) Pilot. The ORD pilot aims to improve and maximize access to and re-use of research data generated by Horizon 2020 projects and considers the need to balance openness and protection of scientific information, commercialization and Intellectual Property Rights (IPR), privacy concerns, security, as well as data management and preservation aspects.

As a participating project, DiSIEM is required to develop a Data Management Plan (DMP), identified as deliverable D8.2. The DMP is a key element of good data management, describing the data management life cycle for the data to be collected, processed and/or generated. The goal is to make research data findable, accessible, interoperable and re-usable (FAIR).

All partners have contributed to the document, completing a project-wide questionnaire that was then used to determine each partner's role in creating and/or processing data.

1.1 Organization of the Document

Since each partner will generate and/or manipulate data, the document is organized with one section per partner (Sections 3-9). Each of these sections is structured in five subsections:

1. **Dataset description** contains a textual description of the dataset. It aims at explaining, in a short paragraph, what the dataset contains and what its goal is;
2. **Standards and metadata** focuses on explaining the internals of the dataset, namely how a user can find syntactical and semantic information;
3. **Data sharing** addresses the issues related to data access, and privacy concerns, namely if the dataset is going to be indexed, and how and to whom it will be made accessible;
4. **Archiving and presentation** covers the aspects related to data availability, during and beyond the project, as well as the actions taken and planned to support availability;
5. **Data details** goes into the specifics of each partner's dataset, describing its content.

Besides these per-partner sections, the document also contains a general description of our overall methodology in terms of data collection and sharing in Section 2. The summary and conclusions of the Data Management Plan are in Section 10. In the appendix, we included the questionnaire each partner filled to prepare the document.

2 Methodology

In this section, we explain some general policy we defined to store and share the data sets produced during the project and the overall methodology used for producing this document.

2.1 DiSIEM Policy for Storage and Sharing of Datasets

One of the most important aspects of the methodology is how datasets are to be stored and used during the project.

A first general concern is how the produced datasets are to be stored. The consortium decided to do that in three ways, for different types of datasets:

- For the public datasets, i.e., the ones we can share outside the consortium, we plan to publish them on the project webpage (or in another public repository to be referred by the project webpage).
- For controlled datasets, i.e., the ones that will be anonymized and shared within the consortium for enabling partners to do exploratory studies, we created a special directory in the project repository for storing them. The idea is to have a subdirectory for each dataset containing not only the dataset files but also a *info.txt* text file with a brief description and metadata of the dataset.
- For privacy-sensitive datasets, i.e., those that contain critical information from partners and therefore require special care in sharing, we decided that partners need to agree on the specifics of how sharing can be done. This might include the signing of specific agreements and protocols between the involved partners. In any case, this should be done between partners, without any direct influence from the consortium.

Regarding the storage of controlled datasets, they will be kept in our project repository, which is maintained in a dedicated KVM virtual machine hosted by FCUL. This VM can only be directly accessed by DI-FCUL system administrators and is externally visible only through the gitlab web interface and through the git protocol over SSL/TLS. All accesses require authentication using valid credentials and access control is enforced. Therefore, we believe an adequate level of protection is provided for these datasets.

As will be clear in the next sections, the preferred formats for datasets are CSV (Comma Separated Values, as specified in RFC4180 [1]) and JSON, since both are text-based and easily parsed by any tool or service being used within the project.

2.2 Data Collection Methodology

To compile the data management plan, a questionnaire was first elaborated covering the main questions that need to be answered in the template provided by the European Commission [2].

In the second phase, each project partner responded to the questionnaire, filling it with as much detail as possible at this stage of the project. Completed questionnaires were stored for analysis and traceability in the project's git repository.

In the third phase, the Data Management Plan was created as a synthesis of the questionnaire results, attempting to take advantage of commonalities between responses to provide a simple view of data management procedures within the consortium.

Further revisions of the document will be based on updates to partner questionnaires. Therefore, the DMP will be updated at least by the mid-term and final reports to be able to accommodate any new data forms and requirements that cannot be estimated in this current stage of the project.

3 Dataset FFCUL

FFCUL is an academic partner in the project therefore it is not expected to contribute with datasets about monitored infrastructures. However, it plans to contribute with some OSINT datasets that might be useful for evaluating the tools and techniques proposed for processing such kinds of data.

3.1 Dataset Description

In principle, FFCUL will provide a collection of tweets classified as "relevant or not" for a given reference infrastructure, a list of operating systems vulnerabilities collected from NVD and enriched with information from other databases, and a list of compromised IP addresses collected from several security feeds on the Internet.

3.2 Standards and metadata

The dataset will contain data formatted using the common Comma-Separated Values (CSV) standard.

3.3 Data Sharing

Since all these datasets are being collected from public feeds from the Internet, FFCUL intends to make them publicly available, respecting possible data protection legislation.

3.4 Archiving and presentation

The dataset will be made available as companion papers exploring them are published. The idea is to have papers using the datasets for validating tools built within the project. Once the papers are made public, the datasets will be made available either through the project webpage or through DI-FCUL webpage.

3.5 Data details

FFCUL will provide three different types of OSINT datasets that can be used to validate different DiSIEM innovations:

- A collection of tweets gathered from 80 cybersecurity-related accounts such as sans_isc, e_kaspersky, alienvault, vuln_lab, etc. These tweets will be manually classified as relevant or not to some synthetic organization infrastructure;
- A list of operating systems vulnerabilities collected from NVD and enriched with information about exploits and patches obtained from other vulnerability databases such as ExploitDB and OSVDB;
- A list of compromised IPs collected from more than a hundred security feeds organized by published date and source.

Notice that “the operating system vulnerabilities” dataset is somewhat similar to the data offered by the vepRisk tool from City (see next section). In the future, we will try to integrate these datasets to avoid duplicating efforts.

4 Dataset CITY

City, being an Academic partner in the project, will be primarily a data consumer rather than a data producer. We plan to analyse the data provided by the project partners to evaluate and test our extensions and plug-ins for diversity and data visualisation.

4.1 Dataset Description

We do plan to also deploy our own testbed to evaluate and test the extensions we build for diversity and probabilistic modelling. The data will consist of synthetically generated network data, as well as data collected from a University honeypot.

We are also building a tool that gathers public data on vulnerabilities, patches and exploits. The tool is made available from the following site (the URL may be updated and change in the future): http://veprisk.city.ac.uk/sample-apps/vepRisk/

4.2 Standards and metadata

The data from our testbed will consist of network traffic, in the *pcap* format, as well as the alerts of the Intrusion Detection Systems (IDS) we will test: Snort, Suricata and Bro. These will be generated in the respective alert format of the tool vendors.

The data from the vepRisk tool can be downloaded from the site in CSV format.

4.3 Data Sharing

Synthetic data from our testbed will be shared with DiSIEM partners without restriction. Data from honeynets, would need to be anonymized first to remove sensitive, confidential and/or private information. Data from vepRisk is available from the public page of the tool.

4.4 Archiving and presentation

The dataset will be disseminated to the consortium via the Git repository.

4.5 Data details

For the vepRisk tool, the data is taken from the public databases on vulnerabilities, patches and exploits and the information on these data are available from the repositories where this data is collected namely, NVD[1], Exploitdb[2] and various patch databases (e.g. Microsoft[3], Ubuntu[4] etc.)

[1] https://nvd.nist.gov/download.cfm
[2] https://www.exploit-db.com/searchsploit/
[3] https://technet.microsoft.com/en-us/security/bulletins.aspx
[4] https://www.ubuntu.com/usn/

Regarding our testbed, we expect the data will include network flows (source and destination IP addresses, source and destination ports, network protocol, timestamp etc.) and the alerts from the IDS platforms.

5 Dataset EDP

5.1 Dataset Description

Having an operating SIEM platform that receives over 10.000 events per second, EDP – Energias de Portugal, SA. has the capability to provide realistic and meaningful data for analysis. The dataset will consist of a significant subset of real events, comprising data from multiple and diverse sources, after adequate pre-processing to ensure that no confidential information is wrongfully distributed.

5.2 Standards and metadata

The dataset will contain data formatted using the common Comma-Separated Values (CSV) standard, as specified in RFC4180 [1].

5.3 Data Sharing

EDP will make data available for the project partners. The specific information to be shared depends on the need presented by the partners, as well as a risk assessment to guarantee legal and business policy compliance. The final dataset details will be indicated in a later release of the DMP.

Information retrieved from EDP's SIEM platform should not be made publicly available due to the critical nature of the data and user privacy concerns.

EDP is investigating tools to enable data masking and/or anonymization. We identified and started performing tests with two of such tools: Python Faker (http://blog.districtdatalabs.com/a-practical-guide-to-anonymizing-datasets-with-python-faker) and ARX (http://arx.deidentifier.org/).

5.4 Archiving and presentation

The dataset will be disseminated to the consortium via the official Git repository.

5.5 Data details

The most relevant SIEM events collected in EDP's platform, with a summary of the respective field set, are presented in the following table.

	Event source					
Field	Firewall	IPS	User authentication	VPN access	Server access	Antivirus
Event name	√	√	√	√	√	√
Source username	X	X	√	√	√	√
Source address	√	√	√	√	√	√
Source port	√	√	X	X	X	X
Source geo country	√	√	X	√	√	X
Destination username	X	X	√	√	√	√
Destination address	X	X	√	√	√	√
Destination port	√	√	X	X	X	X
Destination geo country	√	√	X	√	√	X
Application protocol	√	√	X	X	X	X
File name	X	X	X	X	X	√
Policy name	X	X	X	X	X	√

Table 1 – Data details (EDP)

Field format:

Event name: String (255-character limit);
Source username: String (255-character limit);
Source address: IP Address (IPv4);
Source geo country: String (255-character limit);
Destination username: String (255-character limit);
Destination address: IP Address (IPv4);
Destination port: Integer from 1 to 65535
Destination geo country: String (255-character limit);
Application protocol: String (255-character limit);
File name: String (255-character limit);
Policy name: String (255-character limit).

6 Dataset AMADEUS

6.1 Dataset Description

Amadeus can provide real datasets from different log sources: applications, Firewalls, OS syslog, Antiviruses, Proxy, VPN, IDS, DNS, etc. We need to pre-process and anonymise the data before sharing it with partners.

6.2 Standards and metadata

Two data format will be used for the shared datasets:

1. Comma-Separated Values (CSV);
2. JSON.

A documentation will be provided with each type of dataset to be shared with the partners.

6.3 Data Sharing

Amadeus datasets will be shared with DiSIEM partners depending on the needs presented. However, partners need to ensure that shared datasets should not be made publicly available in any case, due to legal and business policy restrictions.

6.4 Archiving and presentation

The dataset will be disseminated to the consortium via the official Git repository, or any secure file sharing method (in the case of privacy-sensitive data).

6.5 Data details

A summary of the datasets to be shared with DiSIEM partners can be found in the table below:

Source	Description
LSS ASM logs	An administration tool for an authentication and access control management application
HTTP access logs	HTTP logs from an e-commerce application
Cisco, Palo Alto Network	Firewall logs
McAfee	Antivirus
Suricata, Palo Alto, Bro	IDS
Cisco VPN	VPN

Table 2 – Data details (AMADEUS)

The next sections provide a description of the data fields for each dataset.

6.5.1 LSS ASM logs

The logs of an administration tool for an authentication and access control management application. The dataset to be provided is a set of user actions. A user session is a set of user actions with the same session id (PFX, see table below):

Field	Description
PFX	Session id
Orga	Organisation
Action	Type of action performed
userId	User issuing the action
officeId	Office from which the user is connecting
Country	Country Code
IP	IP address
*Browser	Client browser used
*browserEngine	Client browser Engine
*OS	Client operating system

*These fields are derived from the useragent string.

Table 3 – LSS ASM logs (AMADEUS)

6.5.2 HTTP access logs

This dataset will be extracted from a web server of an e-commerce application. The fields are the default HTTP request fields with some additional nested fields extracted from the IP address and the useragent string. More details in the table below:

Field	Description
Datetime	Timestamp
Method	HTTP method
Urlpath	URI path
Status	HTTP status code
http_referrer	HTTP referrer
Useragent	Useragent String
Accespt_language	Accept Language in the HTTP header
Duration	Request processing time
Hostname	Target HTTP hostname
Referrer_uri_proto	Referrer URI protocol
Referrer_hostname	Referrer Hostname
Referrer_uri_path	Referrer URI path
Referrer_params	Referrer Parameters
Ua	Nested Useragent object
remoteclientipaddress	End User or CDN IP address
client_ip	Private IP address of HTTP server
Geoip	Nested Geo coordinates object
isp	Nested ISP object
edge_proxy_cip	End User or CDN IP address
x_forwarded_for	End User or CDN IP address
Jsessionid	The session id of a given request

Table 4 – HTTP access logs (AMADEUS)

6.5.3 Suricata IDS

This dataset is extracted from the Open source IDS/NSM engine Suricata. A brief description of the most relevant fields is provided in the table below:

Field	Description
Category	Threat category
Dest	Destination IP address
Severity	Threat severity
Signature	Threat Signature
Src	Source IP address
Answer	DNS server answer
Date	Timestamp
Dest_nt_host	Destination IP organization
Dest_port	Destination port number
Dns	Nested DNS response object
http	Nested HTTP request object
Eventtype	Suricata event type
Message_type	Request/Reply
Proto	Transport Layer Protocol
Src_nt_host	Same as Dest_nt_host
Ssl_issuer_common_name	SSL certificate issuer name
Ssl_issuer_organization	SSL certificate issuer organization
Ssl_publickkey	SSL certificate public key
Ssl_subject_common_name	SSL subject name
SSL_subject_organization	SSL subject organization
Ssl_version	SSL/TLS version
TLS	Nested TLS requests object

Table 5 – Suricata IDS (AMADEUS)

6.5.4 Cisco Firewall logs

Within the context of a security incident, administrators can use cisco syslog messages to understand communication relationships, timing, and, in some cases, the attacker's motives and/or tools.

Field	Description
acl	Access control list
Action	The status of the actions (e.g. allowed, blocked etc.)
Cisco_ASA_action	The status of the Cisco Adaptive Security Appliance (e.g. allowed, blocked etc.)
Cisco_ASA_message_id	The id of the Cisco message
Description	The Description of the firewall event
Dest_category	The category destination of the event
Dest_dns	Destination DNS
Dest_mac	The physical address of the mac destination
Dest_nt_host	Destination network host
Dest_port	The port destination

Dest_zone	The server destination of the event
Eventtype	The type of the event
Group	The group of servers
Message_id	the ID of the message
Rule_name	The name of the rule
Severity_level	The severity level of the rule

Table 6 – Cisco firewall logs (AMADEUS)

6.5.5 Next-Generation Firewall – Palo Alto Networks (PAN)

This next-generation firewall classifies all traffic, including encrypted traffic, based on application, application function, user and content.

Field	Description
Action	The action taken by the IDS
Application	The application on which the alert was raised
Client_ip	The IP of the client
Client_location	The location of the client
Date	Timestamp
Dest_asset_id	The asset destination ID
Dest_dns	The dns of the destination
Dest_interface	The destination network interface
Dest_ip	The IP of the destination
Dest_zone	The zone of the destination
Dest_nt_host	Destination network host
Eventtype	The type of event (e.g. allowed, blocked etc.)
dstPort	Destination port
Protocol	The communication protocol being used
RuleName	The name of the rule
Server_IP	The IP of the server

Table 7 – Palo Alto Networks (AMADEUS)

6.5.6 Palo Alto IDS

This dataset is also extracted from Palo Alto Networks next-generation firewalls. It contains the events tagged as threats. A description of the most relevant fields is provided below:

Field	Description
Action	Action taken by the IDS
Application	The application that raised the alert
Category	Category of the intrusion
Client_ip	Client local IP address
Client_location	Location of the client in the network
Date	timestamp
Dest_ip	Destination IP address
Dest_hostname	Destination hostname
Dest_interface	Destination network interface

Dest_nt_host	Destination IP organization
Dest_port	Destination Port number
DestinationZone	Destination network zone
IngressInterface	Ingress network interface
Proto	Transport Layer protocol
Session_id	Communication session id
Severity	Severity level (1 to 5)
Signature	Vulnerability signature
SourceUser	Source Username
Src_bunit	Source user business unit
Src_category	Source category
Src_dns	Source DNS server name
Src_mac	Source MAC address
Src_nt_host	Source IP Organization
Src_owner	Source IP Owner Name
Src_port	Source Port Number
Src_zone	Source IP network zone
Threat:category	Threat category
Threat:name	Threat Name
User	Username
User_watchlist	Boolean, true if User in watch list

Table 8 – Palo Alto IDS (AMADEUS)

6.5.7 McAfee ePO

McAfee ePolicy Orchestrator, a centralized security management software for antiviruses, is the source of this dataset.

Field	**Description**
Action	Action taken by McAfee Antivirus
Category	Threat category
Date	Date
Dest	Office ID
Dest_bunit	Destination business unit
Dest_ip	Destination IP address
Dest_mac	Destination MAC address
Dest_nt_domain	Destination IP network domain
Dest_nt_host	Destination hostname
Dest_owner	Destination User name
Detection_method	Firewall detection method
Devent_description	Firewall event description
File_name	Suspicious filename
Fqdn	Fully Qualified domain name
Is_laptop	Boolean, 1 if Laptop used
Logon_user	Username
Mcafee_epo_os	OS name
Os_build	OS build number
Os_version	OS version

Process	Process name
Product	Component creating the event
Severity	Threat severity level
Severity_id	A number mapped to severity
Src	Source IP address
Src_bunit	Source IP business unit
Src_category	Source IP category
Src_mac	Source MAC address
Src_nt_host	Source IP network zone
Src_owner	Source IP owner name
Src_priority	Same as dest_priority
Threat_handled	Boolean for whether threat is handled
Threat_type	Threat Type
User_email	User email address

Table 9 – McAfee ePO (AMADEUS)

6.5.8 Bro IDS

This dataset is extracted from Bro, an open source network analysis framework. Below is a description of the Bro events fields.

Field	Description
Body	Threat description
Category	Threat category
Date	Timestamp
Dest	Destination IP address
Dest_nt_host	Destination IP network zone
Dest_port	Destination port number
Eventtype	Bro event type
File_desc	Suspicious file
O	Organization
Src	Source IP address
Src_nt_host	Same as dest_nt_host
Src_port	Source Port number
Tag::eventtype	Event type
Uid	User ID

Table 10 – Bro IDS (AMADEUS)

6.5.9 Cisco VPN

This dataset contains events from a Cisco VPN server. A description of the dataset fields is in the summary below.

Field	Description
Assigned_ip	Private IP assigned to the user session
Cisco_ASA_user	Username
Date	Timestamp
Duration	VPN session duration in seconds

Eventtype	Cisco VPN event type
Group	Remote Access Group
IP	User Public IP address
Reason	Connection Lost reason
User_email	User email address
User_identity	Full username
Username	Username

Table 11 – Cisco VPN (AMADEUS)

7 Dataset DigitalMR

DigitalMR works with OSINT and has infrastructure to fetch information to create datasets. We intend to fetch information from security related blogs and tweets for a specific timeline of interest. These datasets will be available during the project.

7.1 Dataset Description

Our data consists of openly available content on the Internet from sources including blogs, forums, news, and social networks like Twitter, Instagram and Facebook. This data is either scraped from the sources using our specially built crawlers or fetched using the built-in API of the data sources such as the ones provided by Twitter and Facebook.

7.2 Standards and metadata

The format of the data is in JSON which is widely supported by several applications and is semi-structured. The size of the data can be up to 5 million posts on the Internet depending on the scope of the project.

7.3 Data Sharing

Given that the content of the data might contain information such as usernames, and privacy laws might vary between countries; it is the responsibility of the user of the dataset to make sure that the applicable legislations are respected.

7.4 Archiving and presentation

The dataset will be shared to the consortium via the official Git repository in JSON and will be available for use by the partners.

7.5 Data details

Some of the common fields in the data include the following:

- Author Username
- Author profile URL
- Post URL
- Parent tweet URL (for twitter content)
- Location
- Content/Post (actual content of the data)
- Date
- Tags (added by DigitalMR)
- Relevance (added by DigitalMR)
- Sentiment (added by DigitalMR)

8 Dataset FRAUNHOFER

Fraunhofer does not plan to produce any dataset during DiSIEM. Instead, data provided from the project partners will be analysed using machine learning and visual analytics methods. This may lead to the development of novel representations of the event data produced by the SIEM platforms, as well as the discovery of user- or session-clusters. These results can be used to develop novel visualization tools for SIEM data.

To represent event sequences, Fraunhofer evaluates the embedding, including the bag-of-words approach, event occurrence frequencies within a given sequence and the TF-IDF-score (term frequency multiplied with the inverse document frequency) of events with respect to a given sequence database. Another approach is to define a similarity measure for sequences. To that extend, Fraunhofer developed an embedding of event types into a metric space, where the distance between events correspond to the co-occurrence frequencies within a given sequence database. These feature representations of event sequences will be used to embed the data in 2D or 3D for visualization, as well as to find clusters of sequences and users and to predict whether a sequence is a potential threat.

9 Dataset ATOS

9.1 Dataset Description

Atos dataset will be generated in a testbed specifically prepared for DiSIEM. The dataset will consist of:

- Events generated by applications or sensors installed in the testbed (e.g. Snort, OSSec, netfilter, JBoss, linux kernel, etc), once normalized to the event format used by the XL-SIEM component;
- Alarms generated by XL-SIEM component.

OSINT data or IoC from external feeds such as AlienVault Open Threat Exchange (*OTX*)[5] could also be used by XL-SIEM in Atos testbed.

Since data will be generated in the testbed, no confidential information will be provided in the dataset.

9.2 Standards and metadata

Currently, data generated in Atos testbed can be provided in two formats:

- Comma-Separated Values (CSV);
- JSON.

No documentation or metadata is provided currently with the dataset. The need for such additional documentation will be analysed for a later release of the DMP.

9.3 Data Sharing

Atos will make data available for the remaining DiSIEM partners. Information retrieved from Atos' SIEM platform should not be made publicly available without previous authorization. The specific information to be shared depends on the needs presented by the partners, as well as a risk assessment to guarantee legal and business policy compliance.

9.4 Archiving and presentation

The dataset will be disseminated to the consortium via the official Git repository. Data generated in Atos testbed can be also shared to DiSIEM partners using *Advanced Message Queuing Protocol* (AMQP) protocol such as RabbitMQ Server.

9.5 Data details

SIEM events to be collected in Atos testbed and the final dataset details will be indicated in a later release of the DMP.

[5] https://otx.alienvault.com/

Some event sources to be considered are:

- Firewall;
- Server access;
- Network Intrusion Detection System.

Currently, SIEM events collected (once normalized by the plugins included in the XL-SIEM agent for each specific data source) have the following fields:

Field	**Description**
Type	Type of plugin: detector or monitor
Date	Date (timestamp) on which the event is received from the sensor
Device	IP address of the XL-SIEM agent generating the event in the normalized format
Plugin_id	Identifier of the data source of event generated
Plugin_sid	Type of event within the data source specified in plugin_id
Protocol	Protocol (TCP, UDP, ICMP…)
Src_ip	IP which the sensor generating the original event identifies as the source of this event
Src_port	Source port
Dst_ip	Ip which the sensor generating the original event identifies as the destination of this event
Dst_port	Destination port
Log	Event data that the specific plugin considers as part of the log and which is not accommodated in the other fields.
Data	Raw event's payload, although the plugin may use this field for anything else.
Userdata1 to Userdata9	Fields defined in the normalized event format to allocate relevant information from the specific event's payload. They can contain any alphanumeric information, and on choosing one or another, the type of display they have in the event viewer will change.
Organization	Identify the organization where the agent is deployed.

Table 12 – Data details (Atos)

10 Summary and Conclusions

The Data Management Plan of DiSIEM describes partners' activity related to datasets. It contains a summary of all the information available as of February 28th, 2017. All (but one) partners intend to create datasets and make them available within the consortium.

With respect to *dataset descriptions*, most of the data manipulated by the DiSIEM project is related to security events collected from SIEM systems and processed using various exploratory methods.

With respect to *standards and metadata*, the most prevalent form of data format is Comma-Separated Values (CSV), a textual description of data that is highly common and widely used in the SIEM and big data communities. This format is very easy to manipulate, particularly adapted to sharing over git (as text files are easily versioned) and is understood by a wide range of tools.

With respect to *sharing*, several partners intend to share the datasets for further research and publication, at least in the academic community. Academic research and innovation is the main objective of the data managed in the DiSIEM project. All partners are aware of data sharing limitations due to privacy concerns and legal obligations. When necessary, information will be anonymized or truncated in compliance with the applicable legislation.

With respect to *archiving and presentation*, partners plan to use internal resources and have them available at the time of writing.

Since it is very early in the project, this document only presents preliminary proposals in terms of sharing, volume and archiving. The project is aware of these aspects and will tackle them by updating the present document during the development of the specifications of the experimentations. Therefore, information in this document is subject to change.

List of Abbreviations

API	Application Programming Interface
CSV	Comma-Separated Values
DMP	Data Management Plan
DNS	Domain Name System
IDS	Intrusion Detection System
IoC	Indicators of Compromise
IPS	Intrusion Prevention System
JSON	Java Script Object Notation
OSINT	Open-source Intelligence
SIEM	Security Information and Event Management
VPN	Virtual Private Network

References

[1] Y. Shafranovich. Comma Format and MIME Type for Comma-Separated Values (CSV) Files. RFC 4180. October 2005. https://tools.ietf.org/html/rfc4180

[2] European Commission. Guidelines for FAIR Data Management in Horizon 2020. Version 3.0. July 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Annex – Questionnaire template

Data Management Plan Questionnaire

The DMP has the objective of defining how data generated in the context of the project should be generated, stored, processed and made available. This definition includes both internal data used inside the scope of the project, limited to the partners, and project outputs that are made public so that other entities can benefit from the investment made by the EU Commission.

Description of Data

Give a brief description of the data, including any existing data or third-party sources that will be used, in each case noting its content, type and coverage. Outline and justify your choice of format and consider the implications of data format and data volumes in terms of storage, backup and access.

Will you generate any type of data? (e.g. raw data from systems, transformed/processed information, research results)	[PLEASE FILL IN]
If yes, what type, format and volume of data?	[PLEASE FILL IN]
Do your chosen formats and software enable sharing and long-term access to the data	[PLEASE FILL IN]
Ate there any existing data that exist/you can reuse (link/information)?	[PLEASE FILL IN]

Data Management

Describe the types of documentation that will accompany the data to help secondary users to understand and reuse it. This should at least include basic details that will help people to find the data, including who created or contributed to the data, its title, date of creation and under what conditions it can be accessed.

Documentation may also include details on the methodology used, analytical and procedural information, definitions of variables, vocabularies, units of measurement, any assumptions made, and the format and file type of the data. Consider how you will capture this information and where it will be recorded. Wherever possible you should identify and use existing community standards.

Who created/contributed/owns the data?	[PLEASE FILL IN]
What is the used methodology?	[PLEASE FILL IN]
What is the data's origin? (e.g. application, system or process)	[PLEASE FILL IN]
For whom/end user is the data useful? (e.g. university, research organization, scientific publication)	[PLEASE FILL IN]
Do you see any possibility to integrate or reuse the data in the future? By whom?	[PLEASE FILL IN]
What information is needed for the data to be read and interpreted in the future?	[PLEASE FILL IN]
How will you capture/create this documentation and metadata?	[PLEASE FILL IN]
What metadata standards will you use and why?	[PLEASE FILL IN]

Research Data Identification

Discoverable: Is the data and associated software produced and/or used in the project discoverable (and readily located) and identifiable by means of a standard identification mechanism? (e.g. Digital Object Identifier).

Accessible: Is the data and associated software produced and/or used in the project accessible? If yes, in what modalities, scope and license models? (e.g. licencing framework for research and education, embargo periods, commercial exploitation).

Assessable and intelligible: Is the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review? (e.g. are the minimal datasets sent together with scientific papers for peer review, is the data provided in a way that judgements can be made about reliability and the competence of those who created them).

Useable beyond the original purpose for which it was collected: Is the data and associated software produced and/or used in the project usable by third parties? If yes, is there a validity for this use, or is the data useful even a long time after its collection? (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists).

Interoperable to specific quality standards: Is the data and associated software produced and/or used in the project interoperable, allowing data exchange between researchers, institutions, organisations and countries? (e.g. adhering to standards for data annotation, data exchange, compliant with available software applications, and allowing re-combinations with different datasets from different origins?)

Is the data discoverable?	[PLEASE FILL IN]
Is the data accessible?	[PLEASE FILL IN]
Is the data assessable and intelligible?	[PLEASE FILL IN]
Is the data usable beyond the original purpose for which it was collected?	[PLEASE FILL IN]
Is the data interoperable to specific quality standards?	[PLEASE FILL IN]

Accessibility – Data sharing, archiving and preservation

Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating the type of repository (institutional, standard repository for the discipline, etc.)
Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

In case the dataset cannot be shared, the reason for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

How will potential users find out about your data?	[PLEASE FILL IN]
With whom will you share the data, and under what conditions?	[PLEASE FILL IN]
Will you share data via a repository, handle requests directly or use another mechanism?	[PLEASE FILL IN]
When will you make the data available?	[PLEASE FILL IN]
Will you pursue getting a persistent identifier for your data?	[PLEASE FILL IN]
What data must be retained/destroyed for contractual, legal, or regulatory purposes?	[PLEASE FILL IN]
How will you decide what other data to	[PLEASE FILL IN]

keep?	
What are the foreseeable research uses for the data?	[PLEASE FILL IN]
How long will the data be retained and preserved?	[PLEASE FILL IN]
Where e.g. in which repository or archive will the data be held?	[PLEASE FILL IN]
What costs if any will your selected data repository or archive charge?	[PLEASE FILL IN]
How will these costs be covered?	[PLEASE FILL IN]
Have you costed in time and effort to prepare the data for sharing/preservation?	[PLEASE FILL IN]
How will the data be secure/What are the security's mechanisms you will use to protect the data?	[PLEASE FILL IN]

Intellectual Property Rights

How will the data be licensed for reuse?	[PLEASE FILL IN]
Are there any restrictions on the reuse of third-party data?	[PLEASE FILL IN]
Will data sharing be postponed/restricted? e.g. to publish or seek patents	[PLEASE FILL IN]
When will the data be licensed for reuse?	[PLEASE FILL IN]