Evaluating the Quantity of Incident-Related Information in an Open Cyber Security Dataset

Benjamin Aziz, John Arthur Lee, and Gulsum Akkuzu

School of Computing
University of Portsmouth
Portsmouth PO1 3HE, UK
benjamin.aziz@port.ac.uk
John.Lee1@myport.ac.uk
gulsum.akkuzu@port.ac.uk

Abstract. Data-driven security has become essential in many organisations in their attempt to tackle Cyber security incidents. However, whilst the dominant approach to data-driven security remains through the mining of private and internal data, there is an increasing trend towards more open data through the sharing of Cyber security information and experience over public and community platforms. However, some questions remain over the quality and quantity of such open data. In this paper, we present the results of a recent case study that considers how feasible it is to answer a common question in Cyber security incident investigations, namely that *"in an incident, who did what to which asset or victim, and with what result and impact"*, for one such open Cyber security database.

Key words: Cyber Security Incidents, Open Datasets, Quantity of Information

1 Introduction

When designing a security system, data and information are imperative to creating the best solution. Most organisations will collect and analyse data from their own systems and use these and the experience, as well as the knowledge of the personnel involved, to design a security strategy. Cyber security practitioners working in an increasingly competitive environment face the challenge of providing security that is focused in the areas that threaten their business the most to reduce crime and loss. Because most organisations have limited resources, especially the smaller ones, the solution needs to be cost effective without exposing the organisations to unwarranted liability or embarrassment. Data-driven security, therefore, can help expose hidden patterns of errant behaviour and offer credible countermeasures and effective controls at such a cost effective level.

This obviously has restrictions, and a broader and deeper knowledge of security threats can mean a more robust and effective security system. At the same time, organisations can benefit from the combined knowledge, experience and competences of a wider community in order to improve their understanding of

the risks that they may be facing. Sharing threat information through a community database enables organisations to do just this. By exploiting this shared knowledge, organisations can make more informed decisions to improve their own defences, threat detection practises and mitigation strategies. By collecting and analysing Cyber threat information from multiple sources, an organisation can also enrich existing information and make it more actionable [8]. In recent times, there has emerged a wealth of open data sources to assist Cyber security strategists in understanding systems and threats against them. Notable examples include SecRepo [11], VERIS/VCDB [18, 19], CERT's Vulnerability Notes Database at Carnegie Mellon University [5], CAIDA [4], UNSW-NB15 [13] and the open datasets from the Los Alamos National Laboratory (LANL) [10]. As a result, data-driven security has emerged as a paradigm that uses such data sources effectively to manage the ever-changing risk landscape organisations operate in.

Nonetheless, a very important question in any data-driven technique is the quality of the dataset under consideration. This question is particularly critical in the context of Cyber security data, as was demonstrated recently for example in [9]. We aim in this paper to study this question through the analysis of the quantity of data in an example Cyber security dataset collected by Verizon [18]. Our approach focuses on identifying and extracting information from such dataset relevant to a fundamental question in Cyber incidents analysis, namely: *"in an incident, who did what to which asset or victim, and with what result and impact"*. Our findings show that while organisations are generally happy to report data for the majority of the elements of this question, they are particularly reluctant to report data related to the impact of Cyber incidents.

The rest of the paper is organised as follows. In Section 2, we discuss related work in literature where the problem of the quality of security data in open datasets has been tackled. In Section 3, we give an overview of the dataset used in our case study, namely the VERIS schema and dataset [18, 19]. In Section 4, we describe the methodology used in extracting the sub-schema relevant to our research question highlighted above. In Section 5, we outline this scheme, and in Section 6, we present the results of our quantity of data analysis. Finally, in Section 7, we conclude the paper and outline directions for future research.

2 Related Work

Open Cyber security datasets, such as the ones we mentioned in the Introduction, are becoming increasingly popular, and as highlighted in [14], there is growing trend in encouraging the generation of such datasets. For example, [6] proposed ID2T, a DIY dataset creation toolkit for Cyber security incident detection. Another example of a data-driven security management approach was described in [3], who proposed a method for integrating security system data with systems engineering principles, in an attempt to increase the effectiveness of security systems being designed and implemented and to balance risk, effectiveness and

cost. The method also proposes that human intelligence should not be ignored as a compliment to technological intelligence.

Whilst a few studies have been conducted to investigate the quality of data in more general datasets (e.g. for Wikidata and DBpedia as in [17] and more general, for any linked open data as in [20]), there is still no significant research effort done to understand the quantity of Cyber security data available. Perhaps the most interesting such research so far has been the work of [16], who proposed semiotic levels as a theoretical basis for the definition of data quality in the context of information systems security. For example, the relationship between data and information is an interpretation-related quality, which would affect security operations (e.g. confidentiality, integrity and availability controls), whereas that between information and knowledge is a usefulness-related quality, which in turn would affect decision-making processes (e.g. when enforcing security policies.) On the other hand, the authors in [7, 15] proposed an approach that improves the sharing of Cyber security information through understanding the requirements and constraints underlying the collaborative system.

The work presented in this paper is based on current research effort by the authors, who in [1] demonstrated how open datasets can be effectively used to extract useful information for predicting features of future incidents. Similarly, in [2], the authors demonstrated how open data can be used to evaluate and reason about XACML [12] security policies based on risk probabilities, which offer a quantitative approach to security. However, none of the above works attempted to provide some understanding of the quality or quantity of data in Cyber security-related datasets or platforms.

3 VERIS: A Schema for Cyber Security Incidents

The Vocabulary for Event Recording and Incident Sharing (VERIS) [18] is a dataset and schema defining a set of metadata and metrics for describing Cyber security incidents. It is currently considered a leading provider of open quality information in the IT security domain and provides a framework that organisations can use to collect and share information on security incidents in a responsible and anonymous manner, with the aim of constructing a ground on which researchers and experts in the IT security industry can cooperate to learn from their knowledge and experiences. The VERIS schema itself consists of five general categories, containing descriptions of security incidents:

- *Incident Tracking*: this category contains general information about the incidents, for example, the source identity, summary of the incident and whether the incident is related to other incidents.
- *Victim demographies*: this category contains information related to the organisation being affected by the incident, for example, its country of operation, number of employees, revenue and industry type.
- *Incident description*: this category contains information related to the question of "who did what to what (or whom) with what result". It is based on the so-

called A4 threat model developed by Verizon and contains descriptions related to the Actors, Assets, Actions and Attributes (A4) of an incident.
– *Discovery and response*: this category contains information related to the incident's timeline, its discovery method, root causes and corrective actions.
– *Impact assessment*: this last category contains information on loss categorisation and estimation and impact rating.

The significance of the VERIS dataset lies in the fact that it is a *community-based* dataset. This means that its data are collected from a wide range of industries and varied over different types and sizes of organisations, therefore providing a rich ground for organisations to learn about the various risks and threats that could exist on a global level. This renders the dataset more widely applicable than datasets that are generated in the context of single organisations. The VERIS dataset, known as VCDB [19], had at the time the work reported in this paper was carried out 7834 recorded incidents between 2010 and 2017, with its schema metadata described by some 2398 elements. Both the VERIS framework as well as the VCDB dataset are initiatives by the Verizon RISK team.

4 Data Extraction Methodology

Before evaluating the quantity of information available in VERIS to answer our research question, we can summarise our methodology used in extracting the relevant data by the diagram shown in Figure 1.

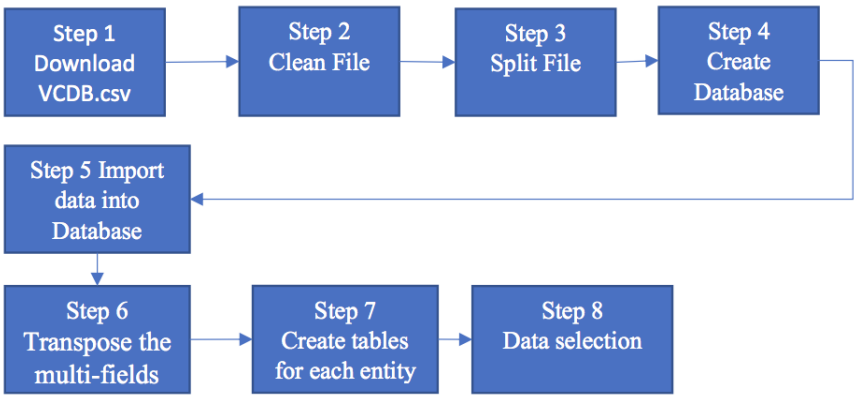


Fig. 1: Our Data Extraction Methodology

This methodology comprises the following eight steps:

1. Download the dataset file, in this case "VCDB.csv". We used the version as of 15 March 2018, which contained 7834 records and 2398 fields, covering years 2010 to 2017.

2. Clean the file, as the file contained many non-standard characters, which had to be removed so that they would not interfere with the loading process. This was done manually in a basic text editor using find and replace.
3. Split the file, since there were 2398 fields in the file, these needed to be split into multiple files in order to make it possible for the data analysis server to process them. This step was performed using a VB.NET script. We used the Microsoft SQLServer Integration Services to perform this step.
4. Create a database, where we used the SQL Server database management studio system to create a database named VCDB, in order to receive the data from the newly created files.
5. Import data into the databse, where each file was imported into separate tables in the new VCDB database, each table having the name of the file from which the data came.
6. Transpose multiple fields, as many of the fields in the dataset are binary data that can be transposed into a single data field. For example, when asking if the ACTOR is of External, Internal or Partner type, the dataset uses 3 fields namely [actor#External], [actor#Internal] or [actor#Partner], each with a binary (i.e. True/False) value. One can instead optimise this into a single field with the three values (Internal, External or Partner).
7. Create tables for each entity, where we focused on the entities selected for the sub-schema. Each table contained also descriptions of the different records.
8. Data selection, which is the final step leading to our extracted sub-schema (described in the next section).

5 The Extracted Sub-Schema

In order to answer the question *"in an incident, who did what to which asset or victim, and with what result and impact"*, we next need to identify the relevant part of the VERIS schema that contains enough information to answer the question. This information can be summarised in terms of the following four categories:

1. Incidents information, which captures some general information related to an incident, such as the various identifiers, date the incident occurred on, levels of confidence and so on.
2. A4 information, which represents information related to the A4 model (i.e. Actors, Actions, Assets and Attributes.)
3. Victims information, which is information related to the victims of the incident, the industry, organisation, revenue and country of the victim.
4. Impact information, which is essential information describing the impact of the incident, e.g. loss type, rating and overall amount.

More specifically, the sub-scheme of VERIS corresponding to these categorites of information is shown in Figure 2. Furthermore, Table 1 defines how the mapping between our research statement concepts and the seven VERIS types captured

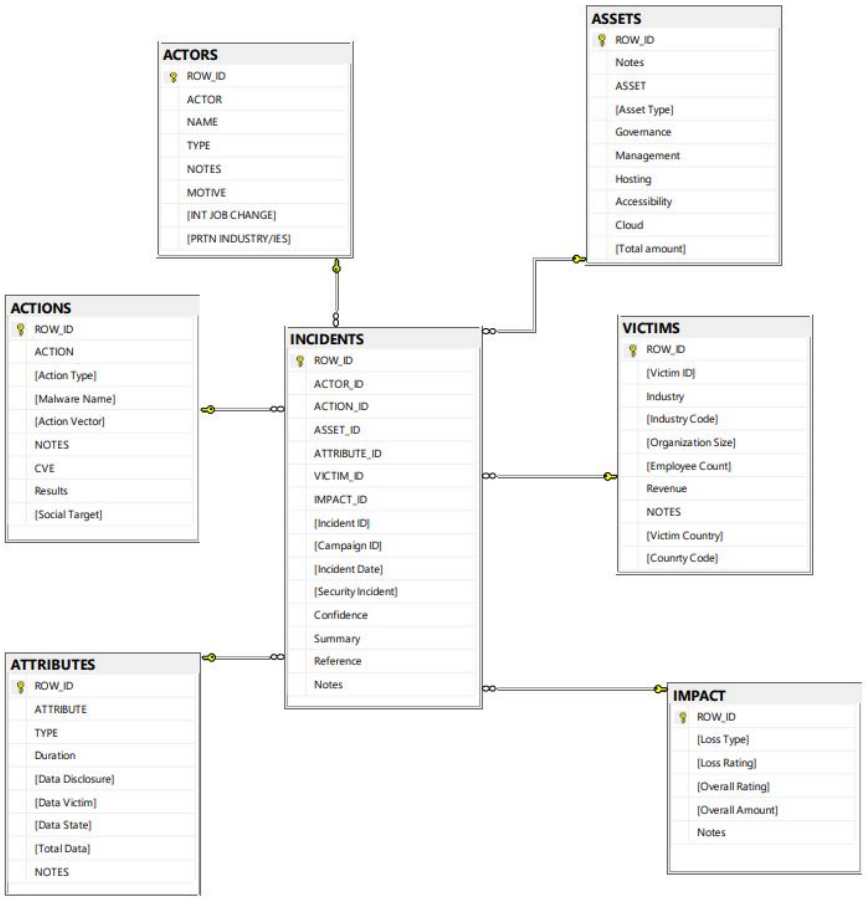


Fig. 2: The extracted sub-schema

in the sub-schema is done, indicating at the same time possible values for some of the sub-types of information in each type.

6 Quantity of Data Analysis

After identifying the relevant categories of information as outlined above in the extracted sub-schema, we next carried out an analysis to measure the amount of information for each element of this sub-schema. After filtering the original dataset of incorrectly inputted records, we ended up with 7210 records. For each of the seven elements of the extracted sub-schema, we measured the percentage of cases that had populated that element over the total number of cases. The results are shown in Tables 2-8 for the extracted schema.

Table 1: Mapping between research statement concepts and VERIS types

Statement Concept	**VERIS Type**	**Possible Values**
Incident	INCIDENTS	Confirmed, Suspect False positive, Near miss
Who	ACTORS	External, Internal, Partner
What	ACTIONS	Hacking, Malware, Misuse Physical, Error, Social Environmental
Asset	ASSETS	Server, Network, User device Media, People, Kiosk/Public Terminal
Victim	VICTIMS	Demographic information
Result	ATTRIBUTES	Confidentiality/Possession Integrity/Authentication Availability/Utility
Impact	IMPACT	Impact assessment information

The first four tables, Tables 2-5, represent the quantity of information available in the A4 (i.e. Actors, Actions, Assets and Attributes) category. We found here that there was an abundance of data; Actors at 97%, Actions at 100%, Assets at 90% and Attributes at 97%. All percentages out of the 7210 records. Therefore, we were able to answer the sub-question *"who did what to which asset and with what result"* for about 90% of the incidents reported correctly.

In terms of the lack of data, we found that for the Actor category, the least populated data items were those related to their identity (10%), sources and capabilities of the actor (38%) and whether they had an internal job change (2%). For the Actions category, we found that the least populated data items were those related to the malware names (2.5%), CVEs exploited by the action (1%), results (3.6%) and social target (2.5%). For the Assets category, we found that the least populated data items were those related to the management and hosting of assets (0%), the accessibility of the assets (0.14%), whether the asset is a Cloud service (0.24%) and what the total amount of the assets was (3.3%). Finally, for the Attributes category, the least populated data item was the duration of the effect of loss or exposure (3%).

Table 6 shows the quantity of information in the VICTIM category. We found that this category of information was well-populated generally, with information supplied for at least 56% of cases, except for information related to the annual revenue of victims (7%), which can be sometimes sensitive information particularly for the case of privately-owned companies. We noticed that the least populated set of data were those belonging to the IMPACT category, shown in Table 7. Most of the fields had fewer than 1% reported data. Finally, for the category of INCIDENTS information, shown in Table 8, the most notable aspect was the hesitance of organisations to report their levels of confidence in the supplied data. Only 8.5% of cases reported any level of confidence.

Table 2: Quantity of ACTORS Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
ACTOR	varchar	Internal, External or Partner	97.00%
NAME	nvarchar	Name or ID of Actor	19.00%
TYPE	varchar	Defines resources and capabilities of capabilities of ACTOR	38.00%
NOTES	nvarchar	Extra information	10.00%
MOTIVE	varchar	Helps to understand intensions	50.00%
INT JOB CHANGE	varchar	Had the employee recently changed job?	2.00%
PRTN INDUSTRY/ IES	varchar	Type of industry of partner US Census NIACS codes	3.50%

Table 3: Quantity of ACTIONS Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
ACTION	varchar	Primary Threat Action	100.00%
TYPE	varchar	What varieties or functions or methods of Primary action were involved	71.00%
Malware Name	nvarchar	Common name or strain of the Malware	2.50%
Action Vector	varchar	What were the vectors or paths of infection or attack	60.00%
NOTES	nvarchar	Enter any additional details deemed noteworthy	8.60%
CVE	nvarchar	Any CVEs exploited by this Action (1)	1.00%
Results	varchar	Exfiltrate, Exfiltrate or elevate	3.60%
Social Target	varchar	Who was the target of these social tactics	2.50%

Table 4: Quantity of ASSETS Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
ASSET	varchar	Asset Category	90.00%
TYPE	varchar	Specific type of asset	93.00%
Notes	nvarchar	Enter any additional details deemed noteworthy	2.20%
Governance	varchar	Who owns / governs the asset	10.00%
Management	varchar	Who manages the asset	0.00%
Hosting	varchar	Where (physically) is the asset hosted	0.00%
Accessibility	varchar	How accessible is the asset	0.14%
Cloud	varchar	If a cloud service what type is it	0.24%
Total amount	nvarchar	Total amount of assets of type affected	3.30%

Table 5: Quantity of ATTRIBUTES Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
ATTRIBUTE	varchar	Confidentiality, Integrity and Availability (CIA)	97.00%
TYPE	nvarchar	What was the nature of integrity/authenticity loss	89.00%
Duration	varchar	Duration of effect of loss or exposure	3.00%
Data Disclosure	varchar	Was non-public data disclosed	89.00%
Data Victim	varchar	Who was the victim within the organisation.	62.70%
Data State	nvarchar	State of data when disclosed	47.00%
Total Data	nvarchar	Number of records affected	56.00%
Notes	nvarchar	Enter any additional details deemed noteworthy	10.60%

Table 6: Quantity of VICTIM Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
VICTIM ID	nvarchar	Identifier or name of victim	96.00%
Industry	nvarchar	Industry	100.00%
Industry Code	float	industry (NAICS code)	100.00%
Organization Size	varchar	Large or small	67.50%
Employee Count	varchar	Number of employees	60.00%
Revenue	nvarchar	Annual revenue of the victim	7.00%
Total Data	nvarchar	Number of records affected	56.00%
Notes	nvarchar	Enter any additional details deemed noteworthy	0.80%
Victim Country	varchar	Country of operation	98.00%
Country Code	varchar	ISO3166-1 two digit country code	98.00%

Table 7: Quantity of IMPACT Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
Loss Type	varchar	Specific category of loss	0.00%
Loss Rating	varchar	Qualitative rating of impact	0.00%
Overall Rating	varchar	Qualitative rating of overall impact	0.35%
Overall Amount	decimal	Most likely estimated money amount	0.83%
Notes	nvarchar	Enter any additional details deemed noteworthy	1.66%

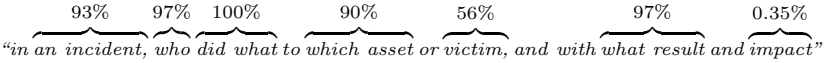
Table 8: Quantity of INCIDENTS Information

Column Name	Data Type	Description	Percentage Populated
$ROW - ID$	float	Unique Row ID	100.00%
$ACTOR - ID$	float	ACTOR Row ID	100.00%
$ACTION - ID$	float	ACTION Row ID	100.00%
$ASSET - ID$	float	ASSET Row ID	100.00%
$ATTRIBUTE - ID$	float	ATTRIBUTE Row ID	100.00%
$VICTIM - ID$	float	$VICTIM - Row - ID$	100.00%
$IMPACT - ID$	float	IMPACT Row ID	100.00%
$Incident - ID$	nvarchar	To uniquely identify incidents for storage and tracking over time	100.00%
Incident Date	date	Date the incident occurred	97.00%
Security Incident	varchar	Was this a confirmed security incident? Confirmed, suspect, false positive or near miss	99.80%
Confidence	varchar	How certain are you that the information you provided about	8.50%
		this incident is accurate? High, Medium, Low or None	
Summary	nvarchar	Brief summary of the incident.	93.40%
Reference	nvarchar	URL or internal ticketing system ID	94.00%
Notes	nvarchar	Enter any additional details deemed noteworthy	5.30%

7 Conclusion

Since data-driven security management effectively helps organisations to understand their situations in terms of Cyber security, extracting knowledge from Cyber security datasets has been a crucial point in recent years towards this understanding. In this paper, we have given a representation of understanding if and to what extent useful free community Cyber security datasets can be to organisations when developing a Cyber security plan. We also showed the possibility of answering the fundamental question, *"in an incident, who did what to which asset or victim, and with what result and impact"*, with a subset of the VCDB dataset. A quantitative analysis was given, which measured the amount of information for each element of the extracted sub-scheme corresponding to the above question.

As a result, one can roughly illustrate quantity of information present in the various parts of the question, as follows:



The significance of these results lies in the wider context of risk analysis. Risk is often defined as the product of the probability of a bad event happening and the

impact of that event. Whilst the amount of information available in answering the majority of the above question helps calculate the probability part of risk, we find that we are quite poorly informed about the impact part.

For future research, we plan to apply more statistical calculations to the VERIS dataset, in particular to measure not just the quantity of information but also its quality. In fact, one important step is to develop open source tools that would automate such evaluations. We also plan to perform similar analyses for other open datasets, such as [4, 5, 11, 13], and also importantly, for proprietary (non-open) data that would be more company-specific.

We consider this kind of research as initial experiments towards a more formal framework for evaluating quantity and quality of open data, where we would define a methodology for performing such evaluations.

References

1. Akkuzu, G., Aziz, B., et al.: Feature analysis on the containment time for cyber security incidents. In: 2018 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR). pp. 262–269. IEEE (2018)
2. Aziz, B.: Towards open data-driven evaluation of access control policies. Computer Standards & Interfaces **56**, 13–26 (2018)
3. Cano, L.A.: A modern approach to security: Using systems engineering and data-driven decision-making. In: 2016 IEEE International Carnahan Conference on Security Technology (ICCST). pp. 1–5 (Oct 2016)
4. Center for Applied Internet Data Analysis: CAIDA Data, `http://www.caida.org/data/overview/`, last accessed: 14.08.2017
5. CERT Coordination Center: CERT Vulnerability Notes Database, `http://www.kb.cert.org/vuls`, last accessed: 14.08.2017
6. Cordero, C.G., Vasilomanolakis, E., Milanov, N., Koch, C., Hausheer, D., Mühlhäuser, M.: Id2t: a diy dataset creation toolkit for intrusion detection systems. In: 2015 IEEE Conference on Communications and Network Security (CNS). pp. 739–740. IEEE (2015)
7. Dandurand, L., Serrano, O.S.: Towards improved cyber security information sharing. In: 2013 5th International Conference on Cyber Conflict (CYCON 2013). pp. 1–16 (June 2013)
8. Johnson, C.S., Badger, M.L., Waltermire, D.A., Snyder, J., Skorupka, C.: Guide to Cyber Threat Information Sharing. Tech. Rep. 800-150, NIST (2016)
9. Liang, G., Weller, S.R., Zhao, J., Luo, F., Dong, Z.Y.: The 2015 ukraine blackout: Implications for false data injection attacks. IEEE Transactions on Power Systems **32**(4), 3317–3318 (July 2017)
10. Los Alamos National Laboratory: Cyber Security Science Open Data Sets, `http://csr.lanl.gov/data/`, last accessed: 14.08.2017
11. Mike Sconzo: SecRepo.com - Samples of Security Related Data, `http://www.secrepo.com`, last accessed: 14.08.2017
12. Moses, T.: eXtensible Access Control Markup Language (XACML) Version 2.0. OASIS Standard (2005)
13. Moustafa, N., Slay, J.: Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: 2015 Military Communications and Information Systems Conference (MilCIS). pp. 1–6 (Nov 2015)

14. Sangster, B., O'Connor, T., Cook, T., Fanelli, R., Dean, E., Morrell, C., Conti, G.J.: Toward instrumenting network warfare competitions to generate labeled datasets. In: CSET (2009)
15. Serrano, O., Dandurand, L., Brown, S.: On the design of a cyber security data sharing system. In: Proceedings of the 2014 ACM Workshop on Information Sharing & Collaborative Security. pp. 61–69. WISCS '14, ACM, New York, NY, USA (2014)
16. Tejay, G., Dhillon, G., Chin, A.G.: Data quality dimensions for information systems security: A theoretical exposition. In: Working Conference on Integrity and Internal Control in Information Systems. pp. 21–39. Springer (2004)
17. Thakkar, H., Endris, K.M., Gimenez-Garcia, J.M., Debattista, J., Lange, C., Auer, S.: Are linked datasets fit for open-domain question answering? a quality assessment. In: Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics. p. 19. ACM (2016)
18. VERIZON: The Vocabulary for Event Recording and Incident Sharing (VERIS), `http://veriscommunity.net/`, last accessed: 21.11.2016
19. VERIZON: VERIS Community Database, `http://vcdb.org/`, last accessed: 21.11.2016
20. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web **7**(1), 63–93 (2016)