International Journal of Information Security (2021) 20:21–38
https://doi.org/10.1007/s10207-020-00490-y

REGULAR CONTRIBUTION

Measuring and visualizing cyber threat intelligence quality

Daniel Schlette[1] · Fabian Böhm[1] · Marco Caselli[2] · Günther Pernul[1]

Published online: 2 March 2020
© The Author(s) 2020

Abstract

The very raison d'être of cyber threat intelligence (CTI) is to provide meaningful knowledge about cyber security threats. The exchange and collaborative generation of CTI by the means of sharing platforms has proven to be an important aspect of practical application. It is evident to infer that inaccurate, incomplete, or outdated threat intelligence is a major problem as only high-quality CTI can be helpful to detect and defend against cyber attacks. Additionally, while the amount of available CTI is increasing it is not warranted that quality remains unaffected. In conjunction with the increasing number of available CTI, it is thus in the best interest of every stakeholder to be aware of the quality of a CTI artifact. This allows for informed decisions and permits detailed analyses. Our work makes a twofold contribution to the challenge of assessing threat intelligence quality. We first propose a series of relevant quality dimensions and configure metrics to assess the respective dimensions in the context of CTI. In a second step, we showcase the extension of an existing CTI analysis tool to make the quality assessment transparent to security analysts. Furthermore, analysts' subjective perceptions are, where necessary, included in the quality assessment concept.

Keywords Cyber threat intelligence · Threat intelligence sharing · Data quality · Threat intelligence formats · Information security visualization

1 Introduction

The last years have seen the emergence of sharing information about threats, cyber attacks, and incidents by organizations. The urge to join forces in the fight against cyber criminals originates from an ever-increasing number of attacks and the related risks for organizations [1,2]. Not only the number but also the complexity of attacks has increased over the years resulting in successful intrusions with more severe forms of security breaches. For individual organizations, it is an almost impossible task to detect these complex and decentralized attacks on their own. Thus, organizations share their available information about incidents and attacks. This information is referred to as cyber threat intelligence (CTI).

However, investigations show that inaccurate, incomplete, or outdated threat intelligence is an important challenge for collaborating organizations [3,4]. More recently, empirical studies with domain experts emphasize that ensuring CTI quality throughout the collaboration process is crucial for its continuing success [5,6]. The exchange and utilization of meaningful threat intelligence depends on measuring and ensuring its quality. This necessity is strengthened as the quality of shared information is stated to have an impact on the required time to respond to an incident [7].

Additionally, it is important to inform stakeholders about the quality of individual CTI artifacts [5]. This can help analysts to narrow down available information to the intelligence actually requiring their attention. Therefore, analysts can come to better informed decisions how to react to incidents reported within the CTI. The other way around, the domain knowledge of security analysts is a very promising source for the "fitness for use" [8] of a CTI artifact. Including experts into the process of measuring quality of threat intelligence is a starting point to assess contextually depen-

✉ Fabian Böhm
Fabian.Boehm@ur.de

Daniel Schlette
Daniel.Schlette@ur.de

Marco Caselli
marco.caselli@siemens.com

Günther Pernul
Guenther.Pernul@ur.de

[1] University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany

[2] Siemens AG, Otto-Hahn-Ring 6, 81739 Munich, Germany

dent data quality (DQ) dimensions. To leverage the domain knowledge of experts, it is necessary to make the data quality assessment transparent to them. In a further step, users should be allowed to contribute their own perception of threat intelligence quality which increases the trust into both platform and threat intelligence [9].

This work centers on two aspects making a contribution to measuring cyber threat intelligence quality. We present a first approach to assess relevant quality dimensions of a standardized CTI data format. For this purpose, we first derive relevant DQ dimensions for CTI and define metrics which allow to measure these dimensions. The metrics are then configured to the STIX format as they rely on its structure. We further differentiate metrics which can be calculated automatically and metrics where input of domain experts is needed. Thereupon, we extend our previously proposed open-source CTI analysis tool to convey CTI data quality to security analysts. The extension helps to provide an indication about the quality of the CTI artifact at hand. Our extension also demonstrates how security analysts can contribute to CTI quality assessment through an interactive visualization.

The remainder of this work is structured as follows: Sect. 2 gives an overview of related work in the field of cyber threat intelligence data quality. A brief introduction to the STIX 2 format can be found in Sect. 3. This section additionally provides an example to illustrate the format, the concept of CTI sharing, and related quality issues. In Sect. 4, we select and structure relevant DQ dimensions. Metrics for the assessment of these dimensions in the context of the specific format are configured in Sect. 5. In Sect. 6, we propose an extension of the STIX format for CTI quality and a possible approach to communicate this quality to users of a CTI analysis tool. This section also describes interviews we conducted with security experts to gain feedback on the proposed approach. Our article concludes in Sect. 7 with a short summary and possible future research directions.

2 Related work

Although CTI and especially quality of CTI are not yet extensively researched topics in the information security field, some related work has already been conducted. We give a short overview of this work hereinafter.

Dandurand and Serrano [10] are among the first to define requirements for a CTI sharing platform. The requirements for such a platform include some form of quality assurance and the provision of adjustable quality control processes. The authors, however, do not specify quality dimensions or metrics to assess the quality of the CTI in their proposed infrastructure.

In 2014, Serrano et al. [11] point out that there is missing support for quality control and management in existing CTI sharing platforms. The authors propose that organizations should install quality control processes to provide multiple measurable quality values. Although the need for quality assessment is discussed, it is not described how such an assessment could be implemented into a platform.

Sillaber et al. [5] perform a series of focus group interviews and discussions with threat intelligence experts. They derive a number of findings on how data quality dimensions influence threat intelligence. They do not identify fundamentally new data quality issues specific to the CTI area. However, the authors give several recommendations for future research and for possibly relevant data quality dimensions. This work does not propose an explicit approach to measure DQ in the CTI context but rather stays on a generic level.

In their survey investigating threat intelligence, Tounsi et al. [7] specifically call for methods to evaluate the quality of threat intelligence. This also applies to the wider organizational security operations center (SOC) context as low-quality CTI is identified to be a pivotal issue [12]. To the best of our knowledge, there is no respective academic work addressing these open issues. Furthermore, none of the currently available commercial threat intelligence sharing platforms is actively measuring CTI quality [7]. With this work, we aim to take a first step into this direction.

3 Structured threat information expression (STIX)

First, this section gives a brief overview of the STIX format. This is necessary as following sections rely on a fundamental understanding of format specifics. The second part introduces a motivational example which is intended to illustrate the STIX format and basic processes of a CTI sharing platform. This example highlights the importance of evaluating CTI quality in the context of a centralized sharing platform with multiple participants.

3.1 STIX format

We base our approach to assess CTI quality on the STIX 2 data format defined and maintained by the OASIS consortium.[1] According to recent analyses, STIX is the de facto standard used for CTI [13,14]. The successor of this format is called STIX 2. It is likely that STIX 2 will reach a similar popularity throughout the next years as it is the format with the most extensive application scenarios [14]. Therefore, our quality assessment is built upon this promising format. Whenever the term "STIX" is used in the remainder of this work, we actually refer to STIX 2.

[1] https://oasis-open.github.io/cti-documentation/.

STIX is a machine-readable, semi-structured format based on JavaScript Object Notation (JSON)[2] to structure and exchange cyber threat intelligence. The format provides two main object types:

1. STIX Domain Objects (SDOs) describing characteristics of an incident and
2. STIX Relationship Objects (SROs) describing the relationships between those characteristics.

SDOs and SROs contain a number of common attributes which are part of any STIX object and additional attributes specific to the respective object type. Common attributes are IDs or the type of the object, whereas exemplary-specific attributes are the motivation of an attacker or the version identifier of a tool.

The current specification of the format conveys twelve SDO types [15]. These allow to provide a holistic view of a cyber incident including both high-level attribution intelligence (e.g., the associated attack campaign or the threat actor) and low-level information (e.g., the data indicating the attack and exploited vulnerabilities).

There are two types of SROs. The first SRO type allows to connect any two SDOs with an explicit relationship highlighting e.g., the vulnerability exploited by a malware. Both can be modeled as SDOs, whereas the logical connection between them is expressed by an SRO. The second SRO type denotes that a specific SDO has been identified. It connects this SDO with an SRO describing the evidential data for this assumption.

SDOs and SROs relevant for a specific threat or incident can be encapsulated by a report. The SDO for this purpose is the *Report* object which references all, respectively, relevant SDOs and SROs.

3.2 Motivational example

In this section, we describe a fictional CTI sharing platform which is used by critical infrastructure providers (e.g., hospitals, energy operators, etc.) to exchange threat intelligence artifacts. Although the platform and the providers in our example are fictional, there is a number of real-world sharing platforms comparable to the described one. The specific characteristics and operation modes of the platform are not relevant to our example which is why we chose a fictional setting. The main goal of the following explanations is to describe the central idea and necessary processes of a CTI sharing platform.

Starting the example depicted in Fig. 1, we can think of a power plant operating a state-of-the-art security operations center (SOC). At some point in time, the alerting mechanisms of the plant's intrusion detection systems (IDS) indicate an ongoing attack affecting various critical systems. Automated systems start the collection of related information through log file and network traffic analyses. Immediately, security experts start their analysis to protect the plant's cyber systems and to gain as much insight into the attack as possible.

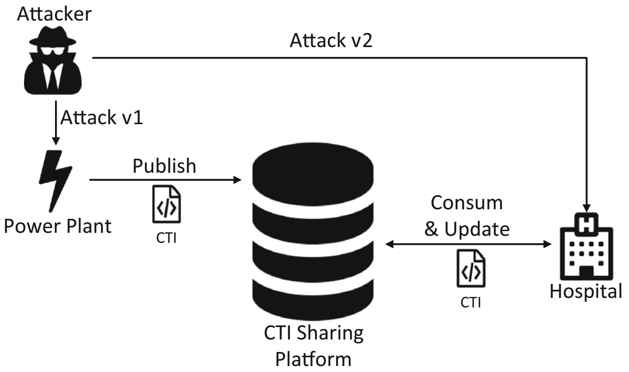


Fig. 1 Simplified CTI Exchange Platform structure

The outcome of automated and manual analyses in the form of collected, attack-related data casts a light on what seems to be an unknown APT. Various machines of the power plant have been compromised and connected to several control units outside of the internal network. The related IP addresses as well as configuration files have been identified. Additionally, the attackers exploited known but unpatched vulnerabilities of a web server and a specific version of an operating system to spread their attack. This allowed them to conduct lateral movement in the organization's network without being noticed. To defend the network and remove the malware, security analysts applied appropriate countermeasures.

Part of the power plant's SOC is the active participation on a CTI sharing platform. On this platform, several operators of critical infrastructure collaborate to improve their cyber defense. Most of these collaborative efforts are based on exchanging intelligence about previously unknown threats or by sharing new insights about existing incidents. There are different roles of participants active on the platform: Publishers post CTI artifacts on the platform, whereas consumers process these artifacts. However, participants of a sharing platform usually hold both these roles simultaneously.

As the power plant's analysts did detect a new type of attack, they transform the gained insights into a STIX report which is published on the sharing platform. The CTI contains the identified threat actor, exploited vulnerabilities, and the deployed malware. Additionally, the analysts include indicators of compromise (file hashes, IP addresses, and the like) to help other participants to detect this attack. They also share the applied countermeasures.

[2] https://www.json.org/.

A simplified example of the STIX artifact shared by the power plant is shown in Listing 1. Please note that some aspects of the example are not fully aligned with the current STIX specification due to readability reasons.[3] However, the example allows to gain a better understanding of STIX. The shared CTI contains the identified *Threat Actor*, the deployed *Malware*, the exploited *Vulnerability*, and an *Indicator* referring to the respective malware file. Additionally, the *Relationships* between these entities are shown. For example, these relationships point out that the *Threat Actor* uses the *Malware* to target a *Vulnerability*.

Another user of the CTI sharing platform might be the operator of a hospital. The operator is leveraging the knowledge made available on the platform to improve the hospital's resilience to cyber attacks. Therefore, published indicators of attacks from the platform are automatically fed into the operator's intrusion detection systems. Additionally, security experts of the operator carry out manual analyses on the most relevant CTI artifacts to identify possible threats. The manual analysis of the artifacts is performed through a visual interface as the CTI format used by the platform is not easily readable for humans.

```
{
  ``type'': ``threat-actor'',
  ``id'': ``threat-actor--1'',
  ``created'': ``2019-04-07T14:22:14Z
     '',
  ``modified'': ``2019-04-07T14:22:14Z
     '',
  ``name'': ``Adversary Bravo'',
  ``description'': ``Is known to
     manipulate critical
     infrastructures, I suppose'',
  ``labels'': [ ``spy'', ``criminal'' ]
},{
  ``type'': ``malware'',
  ``id'': ``malware--1'',
  ``created'': ``2019-04-07T14:22:14z
     '',
  ``modified'': ``2019-04-07T14:22:14Z
     '',
  ``name'': ``Malware d1c6'',
},{
  ``type'': ``vulnerability'',
  ``id'': ``vulnerability--1'',
  ``created'': ``2019-04-07T14:22:14z
     '',
  ``modified'': ``2019-03-07T14:22:14z
     '',
  ``name'': ``A Webserver Vulnerability
     ''
},{
  ``type'': ``indicator'',
  ``id'': ``indicator--1''
  ``created'': ``2019-04-07T14:22:14Z
     '',
  ``modified'': ``2019-04-07T14:22:14Z
     '',
  ``labels'': [``malicious-activity''],
  ``pattern'': ``[ file:hashes.'SHA
     -256' =
    '4bac27393bdd9777ce02453256c5577c
       d02275510b2227f473d03f533924f877
     ']'',
  ``valid_from'': ``2019-04-07T14:22:14
     Z''
},{
  ``type'': ``relationship'',
  ``id'': ``relationship--1'',
  ``created'': ``2019-04-07T14:22:14Z
     '',
  ``modified'': ``2019-04-07T14:22:14Z
     '',
  ``source_ref'': ``threat-actor--1'',
  ``target_ref'': ``malware--1'',
  ``relationship_type'': ``uses''
},{
  ``type'': ``relationship'',
  ``id'': ``relationship--2'',
  ``created'': ``2019-04-07T14:22:14Z
     '',
  ``modified'': ``2019-04-07T14:22:14Z
     '',
  ``source_ref'': ``indicator--1'',
  ``target_ref'': ``malware--1'',
  ``relationship_type'': ``indicates''
},{
  ``type'': ``relationship'',
  ``id'': ``relationship--3'',
  ``created'': ``2019-04-07T14:22:14Z
     '',
  ``modified'': ``2019-04-07T14:22:14Z
     '',
  ``source_ref'': ``malware--1'',
  ``target_ref'': ``vulnerability--2'',
  ``relationship_type'': ``targets''
}
```

Listing 1 Exemplary STIX 2 artifact

The power plant's CTI artifact is analyzed by the hospital's security personnel only a few months after the respective incident. This is mainly because vast amounts of available CTI hinder the security experts to identify threat intelligence relevant for them. During the analysis of the artifact published by the power plant, the responsible security analyst of the hospital spots that the respective attack targets a software in use by the hospital as well. Subsequent network and endpoint analyses indicate that the hospital has been affected although the IDS seems to have not noticed the compromise as the binaries of the malware have changed in the meantime. In addition, although the same software is in use, the version number proclaimed to be exploited at the power plant seems to be invalid.

During the analysis of the incident at the hospital, analysts come across some changes and additional insights into the attack. Additionally, the proposed countermeasures are not sufficient to get rid of the attacker. Therefore, an updated ver-

[3] Object IDs are not in UUIDv4 format, and some mandatory schema structures are left out.

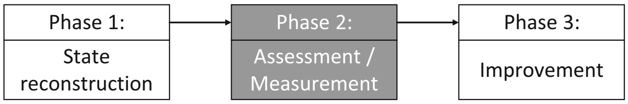


Fig. 2 Process steps of DQ methodologies [16]

sion of the CTI artifact is published to the platform to ensure each participant is informed about the advanced version of the cyber attack. However, during this process the information about the threat actor is unintentionally duplicated leading to redundant information.

The example above shows that the timely exchange of high-quality CTI is crucial for the effort of organizations to prevent cyber security breaches. However, there are numerous pitfalls regarding the quality of the shared threat intelligence. Examples from the above-described use case are: 1) inaccurate information caused by input errors made during the documentation of an attack (invalid version of exploited software), 2) outdated information caused by delays in CTI propagation (changed binaries of malware), or 3) duplicated information caused by collaboration (redundant description of threat actor). Even the overload of CTI available to human analysts and their incapability to determine the most relevant CTI can be seen as a data quality problem. Each of these examples stresses the urge to measure CTI quality and to visualize the results for human analysts.

4 Approach for CTI quality assessment

General DQ methodologies consist of three main process steps depicted in Fig. 2. Initially, the collection of necessary data is performed. Data sources and involved costs are fundamental building blocks for the following process steps. The second step includes the identification and measurement of relevant quality dimensions in the context where the methodology is applied. After quantifying data quality, the last process step strives to improve the quality following a fixed set of techniques and strategies. Although there is no cohesive methodology for information quality management of CTI yet, this work solely focuses on measuring DQ in the context of CTI as highlighted in Fig. 2. Up to now, existing work has mostly provided general advice for mainly the first and the last methodology step but has not described approaches to actually measure CTI quality [5,13,17]. We put explicit focus on the quality assessment. We thus assume the existence and availability of the necessary data for assessment.

Our work on selecting and structuring DQ dimensions relevant for CTI is the result of an iterative process in which we actively sought input and feedback from a number of CTI researchers and practitioners, e.g., domain experts from computer emergency response teams (CERTs). Throughout multiple evaluation iterations the relevant dimensions and their structure as described in the following two subsections were consequently adapted according to the input of the experts.

4.1 Selecting relevant DQ dimensions for CTI

Before introducing measurements for CTI quality, relevant DQ dimensions have to be selected. Extant work has already suggested a wide variety of different DQ dimensions referring to either the data values or the data schema [18]. The literature distinguishes three main approaches for proposing general and abstract quality dimensions: the theoretical approach [19], the empirical approach [20], and the intuitive approach [21].

Considering the different approaches and various DQ dimensions, it is not an easy task to select relevant and applicable dimensions for a problem at hand. Following the empirical approach by Wang and Strong [20], related research such as the work of Sillaber et al. [5], Sauerwein et al. [13], or Umbrich et al. [22] identify a first set of relevant dimensions which is refined throughout this work.

Our resulting set of dimensions is shown in Fig. 3. An interesting finding yielding from the discussion with the CTI experts is the high complexity of the *Appropriate amount of data* quality dimension. This dimension is meant to help experts to decide whether a CTI artifact by any chance could contain helpful information. In general, this decision can only be made by comparing the real-world artifact with its CTI description. However, this is rarely possible. Therefore, another approach is needed to give security analysts an indication for this dimension. Throughout our discussions, it turned out that experts are often basing their decision on the diversity of SDO types and their interconnection in a STIX report. Arguably, homogeneous SDO types and few relationships between them lead to experts' perception that the report does not describe the real-world incident properly. For the in-depth examination of the *Appropriate amount of data* quality dimension we refer to Sect. 5.3.

4.2 Structuring DQ dimensions for CTI

Our goal for DQ assessment in the context of CTI is to come up with measures to quantify the selected dimensions and aggregate them into a combined score for a STIX report. We therefore structure the dimensions in three different levels depending on the input data as shown in Fig. 3. The assessment of the dimensions on the "Attribute Level" operates on specific attributes of STIX objects, e.g., the dimension of *Timeliness* can be assessed using the *modified*- and *created*-attributes of STIX objects. The two dimensions located on the "Object Level" in Fig. 3 are not bound to predefined attributes

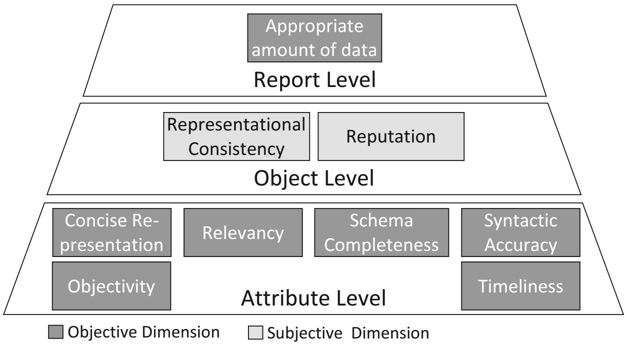


Fig. 3 Schematic of the structure of DQ dimensions

of the objects. In fact, they can be measured based on either varying attribute sets (*Representational Consistency*) or the object as a whole (*Reputation*). At the highest level ("Report Level"), we propose a final dimension to cope with experts' requirement to be informed about whether a report is likely to contain an *Appropriate amount of data* as described in the paragraph above.

Individual scores on both attribute and object level are then aggregated to a combined object quality indicator. This aggregation provides a quick and helpful insight for any user navigating through cyber threat information. Artifacts with a high-quality score are probably the ones to analyze first. Additionally, on a "Report Level" this aggregation allows to inform users about the average object quality in a given report. This is accompanied by an indication whether the report contains an appropriate amount of data. However, as DQ dimensions can be of varying importance for different users the aggregation has to be customizable [11]. Adjustable aggregation parameters enable CTI users to define the weight of each DQ dimensions in the procedure of calculating a quality indication for each STIX object. The corresponding metrics for aggregation are further outlined in Sect. 5.4.

Additionally, to these various levels for the DQ dimensions, we differentiate objective and subjective dimensions which are also indicated in Fig. 3. DQ has to be evaluated with objective measurements as well as from subjective perception [16,23,24]. Objective measurements rely on mathematical and quantitative measures to evaluate an artifact's quality. However, some dimensions of DQ are dependent on their contextual environment. It is thus necessary to incorporate the requirements, experience, and knowledge of domain experts. When it comes to the decision whether data is of high quality regarding a specific use case or context, objective DQ dimensions fail to provide reasonable quality scores [25]. At this point, it is necessary to incorporate subjective measures as a supporting concept. Here, the assessment of an artifact's quality is based on qualitative evaluation by data administrators and other experts.

In the context of a CTI sharing platform, the concept of subjective perception and domain knowledge to evaluate various DQ dimensions equally applies. While domain knowledge is a necessary input for subjective quality dimensions, it also supports assessment of objective DQ dimensions. The domain knowledge can be captured through a system similar to a reputation system where users provide their perception about the quality of an object or report [26]. The need for a reputation system to include subjective quality perceptions and to increase trust is also highlighted in empirical studies [5]. Subjective quality assessment in the CTI sharing context can originate from different stakeholders of a respective platform: On the one hand, consumers (security experts, analysts, etc.) contribute with their domain knowledge and their organization-specific background; on the other hand, a platform host can act as a trusted third party contributing to the quality assessment.

Overall, these three levels provide good and transparent indicators for the quality of a STIX-based cyber threat intelligence artifact. For indication of individual DQ dimensions, we adopt and extend existing naming conventions [20].

5 Measuring CTI quality

In this section, we elaborate on suitable DQ dimensions as the result of our studies. For each dimension, its applicability to the CTI context is described and respective metrics for assessment are configured. Those assessments are either of an objective or a subjective nature depending on whether they can be automated or need manual input. Subjective metrics are based on the perceptions and expressions of a CTI sharing platform's participants. Furthermore, there is a number of objective dimensions which benefit from additional manual input of domain experts. The ordering of the metrics follows the previously outlined structure of the dimensions in Sect. 4.2.

The proposed metrics in the following are again the result of an iterative process collaborating with CTI researchers and practitioners. Several metric configurations result from long discussions with domain experts where a lot of very valuable feedback was provided highlighting possible configurations to assess CTI quality.

Configurations for the metrics are based on the formal ground truth defined in Eqs. 1–5. We formally define two different attribute sets of STIX as A_r (i.e., Eq. 1) and A_o (i.e., Eq. 2). Required attributes a_r, for example, are unique IDs, names, labels, and types which are present in most STIX objects. As for optional attributes a_o, characteristics such as descriptions, versions, and external references are referred to Eq. 3 which defines any STIX Domain Object or STIX Relationship Object as a specific subset of both the available required and optional attributes. This subsequently allows us to describe the objects O held by a CTI sharing platform as a set of objects where each object o is either a SDO or SRO (i.e., Eq. 4). STIX objects such as *Threat Actor*, *Malware*, or *Indicator* belong to the set of SDOs, while *Relationship* and *Sighting* objects are SROs. When an incident or an attack is reported to the platform, the resulting report r is defined by Eq. 5 to be a subset of all objects persisted in the platform.

$$A_r = \{a_r \mid a_r \text{ required in STIX 2}\} \tag{1}$$

$$A_o = \{a_o \mid a_o \text{ optional in STIX 2}\} \tag{2}$$

$$SDO,\ SRO \subseteq (A_r \cup A_o) \tag{3}$$

$$O = \{o \mid o \in (SDO \cup SRO)\} \tag{4}$$

$$R = \{r \subseteq O\} \tag{5}$$

5.1 Attribute level

This subsection defines the DQ dimensions we consider to be assessed at the attribute level, meaning that they rely on a subset of a STIX object's attributes.

Concise representation Concise representation addresses expressiveness of CTI and redundancies within the data [20]. Intensional and extensional are two distinct forms of conciseness. While the former is centered on the uniqueness of attributes and the schema level, the later emphasizes on unique objects. In the motivational example, the duplication of the information about the attacker links to the concise representation dimension as DQ is affected. It is worth noting that in the extant literature, concise representation sometimes only refers to compactly represented information [18]. In the context of STIX, the specification provides clear guidance how to implement a concise representation. It is explicitly stated that a unique identifier is assigned to each artifact. Additionally, each STIX object adheres to a specified JSON schema, and thus, optional and mandatory attributes are predefined. In general, the assumption holds that intensional conciseness is warranted through the schema definition. One exception in STIX is based on specifics of several STIX objects[4] as they contain lists referencing other objects. These lists are prone to redundant inputs, especially when defined manually.

With regard to extensional conciseness, the information within a CTI platform must be assessed for its respective quality. The main reason for this is that with a growing number of CTI producers, the probability of duplicated objects within the platform becomes likely. More precisely, there is a high chance that two or more objects on the platform are semantic duplicates. Even considering one single STIX report, semantically unique objects are not guaranteed as more than one person could work on the documentation of the incident and already existing information might be overlooked. Especially, when taking a look at the numerous free-text description fields defined in the current STIX specification, an indication whether these descriptions contain redundant information is important. However, comparing text for semantic redundancy is not an easy task. We encourage the application of methods for semantic similarity. The *Simhash* algorithm is one example proposed to approach this problem [27]. It allows for comparing two STIX objects regarding their uniqueness. An object o_1 is considered unique in a set of objects O if its $similarity$ to any other object $o_2 \in O$ is below a threshold t (see Eq. 6).

Objective metrics alone are not sufficient to assess concise representation in practical use. It is inevitable to include subjective perceptions through the utilization of domain knowledge. In this case, platform users conduct or support quality assessment and contribute by pointing out redundancies.

$$CR(o) = \begin{cases} 1 \text{ if } similarity(o_1, o_2) < t \\ 0 \text{ else} \end{cases} \tag{6}$$

Objectivity CTI is oftentimes created by multiple human actors during the analysis of an attack. These human CTI creators contribute not only objective threat information but might also introduce emotional or subjective perceptions. Most of the resulting descriptions are phrased in natural language. This is also the case in the motivational example in Sect. 3.2 and the threat actor description. There, the words "I suppose" indicate subjectivity and the context-depended observations of the security analyst. However, objectivity is a desirable characteristic of shared CTI artifacts as only objective information can be helpful for others. Natural language processing and sentiment analysis, therefore, can facilitate the assessment of unbiased and impartial CTI information as part of the objectivity DQ dimension.

Subjective descriptions of CTI information can be identified through the use of various subjectivity detection methods [28]. In the context of CTI and with regard to STIX, special focus is on attributes with free-text description fields in contrast to predefined enumerations and open vocabularies. This ultimately leads toward a sentence-level orientation for subjectivity detection as these fields contain only a limited number of words. Subjectivity detection methods in general can follow a syntactical approach or center on semantics. A thorough investigation into specifics of such methods must be considered during implementation to determine the best-fitting approach. Regardless of implementation, we classify relevant attribute values $v(a)$ of STIX objects into two distinct categories objective and subjective as shown in Eq. 7. Underlying this classification is the application of a suitable sentiment algorithm which yields a score for either objectivity or subjectivity. The results of the classification for chosen

[4] Examples are the *Report* object as well as the *Sighting* object.

attribute values are then aggregated to provide an objectivity metric for each object o based on Eq. 8.

$$OB(a) = \begin{cases} 1 \text{ if } v(a) \text{ classified as } \textit{objective} \\ 0 \text{ if } v(a) \text{ classified as } \textit{subjective} \end{cases} \tag{7}$$

$$OB(o) = \frac{\sum_{a \in o} OB(a)}{|o \cap (A_r \cup A_o)|} \tag{8}$$

Relevancy Relevancy forms a DQ dimension incorporating a user's perspective by comparing sets of property values to assess the usefulness of a CTI artifact for the consumer. This is an important aspect of CTI's fitness for use regarding an individual organization or analyst. For example, CTI describing an incident targeted at a specific industry sector is likely to be less relevant for other industry sectors. Also, security analysts might not be interested in threats targeting technologies not deployed in their organization. To illustrate this, the motivational example hints at the exploitation of vulnerabilities in software used at both the power plant and the hospital. Information about the relevance can be very helpful for analysts when prioritizing CTI artifacts to be analyzed.

Contextual information about the user can either be collected by the platform host or can be found in STIX objects describing the user. Specific characteristics (e.g., the industry sector) of a CTI publisher and those of a consumer are assessed for matches. In addition, attribute values for available STIX objects—for example, the *Vulnerability*—can be compared with the user's characteristics (e.g., the applied technologies), too. The coverage ratio expressed in Eq. 9 indicates relevance by taking the sets of all property values for consumer PV_c, publisher PV_p and relevant STIX objects PV_o into consideration. Congruent property values are set in relation to the total number of property values available for comparison.

The metric for the DQ dimension of relevancy could be further extended by inclusion of information contained in STIX *Sighting* objects. These objects incorporate a number describing how many times the referenced object has been identified. Therefore, this fosters the assessment of relevancy as frequently seen objects (e.g., an *Malware* object) might indicate a high relevance of these objects. This assumption can be expressed in a weighting factor added to the general metric and thus improve DQ assessment.

$$RE(o) = \frac{|PV_c \cap (PV_p \cup PV_o)|}{|PV_p \cup PV_o|} \tag{9}$$

Schema completeness The general completeness of data is confined to the assessment of schema completeness in the context of CTI. To distinguish this data quality dimension from syntactic accuracy, we focus on optional attributes and their values as the STIX JSON schemes already allow to assess the existence of required attributes. This aspect is covered by the DQ dimension of syntactic accuracy later on.

STIX-based threat intelligence can be assessed for schema completeness of individual optional attributes a_o. A missing optional attribute value $v(a_o)$ is identified and classified according to Eq. 10. A strict distinction between complete (i.e., with value) and incomplete (i.e., without value) attributes is enforced. Referring to the example in Sect. 3.2, the vulnerability could be described in more detail with an external reference to a specific Common Vulnerabilities and Exposures (CVE) entry. This optional information would help others to gain further information about the actual vulnerability, how it is exploited, and how it can be fixed. This would ultimately improve CTI quality significantly by making it easier for others to leverage the CTI. In a second step, schema completeness for an entire STIX object o builds upon the previously calculated completeness scores for included attributes. The ratio of filled optional attributes to the total number of optional attributes of an object represents the schema completeness metric as shown in Eq. 11.

$$SC(a_o) = \begin{cases} 1 \text{ if } v(a_o) \neq NULL \\ 0 \text{ else} \end{cases} \tag{10}$$

$$SC(o) = \frac{\sum_{a_o \in (o \cap A_o)} SC(a_o)}{|o \cap A_o|} \tag{11}$$

Syntactic accuracy The data quality dimension of accuracy contributes to the correctness of data. With focus on syntactic accuracy in the context of CTI, the data schema is of particular importance for quality assessment. Syntactic accuracy gives a first indication on the extend to which an object is aligned with its data format.

The OASIS consortium behind the STIX format provides a JSON schema for each object. This allows for an automated matching of objects against those schemes to assess syntactic accuracy. In general, this DQ dimension is measured based on the analysis of attribute values $v(a)$ with $a \in (A_r \cup A_o)$ being part of a domain D [16]. In application to STIX-based threat intelligence, we can use the existing JSON schemes and validate each attribute value against the schema definition. The domain D is derived from the JSON schema which provides data types and allowed values. The assessment for syntactic accuracy of each attribute value is expressed by Eq. 12. An overarching indicator for syntactic accuracy of an object o can, respectively, be calculated as shown in Eq. 13.

$$SA(a) = \begin{cases} 1 \text{ if } v(a) \in D \\ 0 \text{ else} \end{cases} \tag{12}$$

$$SA(o) = \frac{\sum_{a \in o} SA(a)}{|o \cap (A_r \cup A_o)|} \tag{13}$$

Timeliness In the context of CTI, time ascends to one of the crucial elements of CTI quality. As stated earlier, outdated intelligence is identified throughout the relevant literature as one of the core challenges [3,5,13]. It is quite evident that the most current and up-to-date CTI artifacts probably implicate the most value for any type of analysis.

Time-based information contained within CTI data builds the basis for the configuration of a timeliness metric applicable to the CTI context. In general, various metrics can be utilized to assess timeliness. Considering the STIX data format, a basic timeliness metric is described in Eq. 14. The two components of this metric—currency and volatility—are present in every STIX object or can be derived from inherent features of the CTI platform. Volatility in this setting is expressed by the number of modifications to the assessed STIX object. The number of modifications can be drawn if concepts like the historization from earlier work are implemented [29]. This concept allows to track changes and the number of changes applied to a STIX object. Currency is referring to the age of the information and thus the time since its last modification. However, this metric entails certain problems specifically with regard to interpretability as well as to other requirements [30].

Where statistical data about the decline of timeliness for specific CTI information does exist, the metric for timeliness must be adapted. Resulting values of a statistical timeliness metric shown in Eq. 15 can subsequently be interpreted as probability of up-to-date CTI information. Considering the example in Sect. 3.2, the decline for certain STIX objects is higher than for others. File hashes as in the *Indicator* of Listing 1 will likely have high decline values as, for example, malware binaries might undergo slight changes frequently leading to changed hash values. In contrast, information regarding the threat actor might not change in time, thus having no statistical decline at all.

In contrast to these metrics, specific assessment of STIX-based CTI for the DQ dimension of timeliness can also be based on characteristics of STIX objects. For example, *Sighting* objects can provide information about the time of occurrence of referenced STIX objects. It can be thus inferred that for the timeliness of referenced STIX objects, the concept of inheritance applies. STIX objects of type *Observed Data* can be assessed for timeliness following the same procedure. Our proposed metric described in Eq. 16 includes the current time, the time of last occurrence, and a predefined time-based threshold value to foster the applicability of timeliness to any given CTI use case. In general, we focus on objective metrics of timeliness. Subjective perceptions such as expert knowledge about threshold values assist the assessment and can be considered further during implementation. Referring back to the motivational example, the hospital's security analysts can define a threshold based on their experience that indicators are outdated after a specific amount of time.

$$TI_{Basic}(o) = \frac{1}{(Currency(o) \times Volatility(o)) + 1} \tag{14}$$

$$TI_{Statistical}(o) = \exp\left(-Decline(o) \times Currency(o)\right) \tag{15}$$

$$TI_{Assisted}(o) = \begin{cases} 1 \text{ if } t_{current} - t_{last} < threshold \\ 0 \text{ else} \end{cases} \tag{16}$$

5.2 Object level

On the object level, we consider two dimensions which rely on manual input and are therefore defined to be subjective dimensions. They center on object characteristics of a higher abstraction level and often follow a cross-object perspective.

Representational consistency In general, the assessment of representational consistency relies on a set of rules C and semantic conditions c_j contained therein for the underlying data [24]. This DQ dimension needs to be adjusted to the requirements of the individual context and the given use case. Analogous to schema completeness, representational consistency goes beyond aspects of syntactic accuracy. For the context of threat intelligence, representational consistency allows for the enforcement of additional formal requirements which are not addressed by the dimensions of syntactic accuracy or concise representation. These might originate from data format requirements or requirements imposed by a CTI sharing platform. In the following, we propose two exemplary conditions configured to the STIX data format. CTI platforms could define further conditions or adjust existing ones. This is part of an iterative approach to support an increasingly detailed assessment of representational consistency.

In the context of STIX-based threat intelligence, we suggest a first condition to represent the necessity of existence of referenced STIX objects. For all STIX objects, the following "inter-relation constraint" [16] applies: referenced objects of embedded relationships must exist. Moreover, considering individual STIX objects specific relationships must be verified. This applies for all SROs as they connect per definition two SDOs. A second exemplary condition takes time-based information and the chronological order of creation and modification of CTI into account. Hence, it must be verified on the "intra-relation constraint" level that the creation time of any object is prior or equal to the time of modification. Besides, SROs can connect two SDOs only after their creation. Creation time of the corresponding SDOs must be prior or equal to creation time of the SRO. Listing 1 reveals those two exemplary conditions for representational consistency, too. For the *Vulnerability*, modification time precedes creation time by a month. With regard to referenced objects, a *Relationship* (i.e., "relationship–3") points toward a nonexistent *Vulnerability* (i.e., "vulnerability–2").

The assessment of representational consistency on a condition basis is described in Eq. 17. A given STIX object is assessed for each defined condition $c_j \in C$ separately, and the results indicate if a condition is fulfilled. Representational consistency per object is aggregated over all defined conditions in the set of conditions C as seen in Eq. 18.

Please note that although the assessment of an object o regarding a condition c_j can be automated and therefore is objective, the definition of the respective conditions is fully in control of the responsible stakeholder. Thus, we interpret this dimension to rather be subjective than objective with respect to the definitions in Sect. 4.1.

$$c_j(o) = \begin{cases} 1 \text{ if } o \text{ fulfills condition } c_j \\ 0 \text{ else} \end{cases} \tag{17}$$

$$RC(o) = \prod_{j=1}^{|C|} c_j(o) \tag{18}$$

Reputation It is important to build trust in shared CTI environments. Trust and the assessment of trustworthiness can build upon the DQ dimensions of reputation, provenance, and believability. The introduction of two quality sub-dimensions for reputation—reputation of the publisher (i.e., provenance) and reputation of the data set (i.e., believability)—allows for a holistic coverage of the trustworthiness concept in the context of CTI exchange. Our proposed assessment is based on functionalities similar to reputation systems and external human input. Reputation scores for a given publisher p might adhere to a five-star rating system as shown in Eq. 19 as well as reputation scores of a STIX object o as shown in Eq. 20. Based on these reputation scores s contained in a set of scores S, an overall reputation $RS(x)$ for either publisher or STIX object is calculated according to a simple ratio function described in Eq. 21. Sample size $|S|$ supports data quality assessment further and constitutes a relevant additional data point. In the situational example, the hospital can articulate trust toward the power plant and its CTI by rating them accordingly.

While the above-mentioned configuration of reputation is purely subjective, possibilities exist to assist the quality assessment with objective metrics. For one, a list of trusted CTI publishers can be introduced as an indicator for the reputation of a publisher. An analogous indication for the reputation of an object is the number of access requests to a certain artifact set in relation to the number of CTI platform consumers having taken remediating steps upon the threat intelligence.

$$RS(p) = \{s \mid 1 \leq s \leq 5 \ \wedge \ s \in \mathbb{N}\} \tag{19}$$

$$RS(o) = \{s \mid 1 \leq s \leq 5 \ \wedge \ s \in \mathbb{N}\} \tag{20}$$

$$RS(x) = \frac{\sum_{s \in S} s_i}{|S|} \tag{21}$$

5.3 Report level

On the upmost level of Fig. 3, we place a single dimension which takes a complete STIX report including its contained SDOs and SROs into consideration.

Appropriate amount of data The requirement to include the appropriate amount of data quality dimension arose during our discussions with domain experts as described earlier. However, the application of a generic metric proves not feasible due to its semantic component in the form of needed data units. We therefore base our metric on the additional comments of security analysts. Homogeneous SDO types and very few relationships seemingly lead to the experts' perception that the report in general is not very helpful.

To distinguish between a report with homogeneous STIX objects and one with rather diverse objects is a matter of implementation and cannot easily be compressed into a metric. As described above, this is a rather complex task which needs further research efforts. As a first approach toward a feasible support of security experts, we propose a clear representation of occurrences of each STIX object in an artifact. This is achieved by simply counting the instances of the different SDO types within a report. Visualization can provide this relevant information at the report level and can aid DQ assessment at first glance.

Besides this, we take graph theory for the connectedness of the STIX report's SDOs into account. We argue that a metric based on the number of relationships can provide a basic indicator to assess this DQ dimension. In general, the metric for the DQ dimension of appropriate amount of data should yield a higher score for CTI which is densely connected. A given STIX report depicts a graph, and its contained SDOs represent vertices. SDOs are furthermore connected with each other through SROs which resemble edges from a graph perspective. The metric in Eq. 22 sets the number of existing SROs in a given STIX report in relation to the maximum possible number of SROs as defined by the number of SDOs for this report.

The metric for the appropriate amount of data is a challenge for future work. Our simplistic metric could be improved in different ways. A possible direction is a statistical comparison of all available reports. Calculating a report's score for the diversity of SDOs and the respective relationships as a comparison with a baseline diversity from other reports might be a feasible direction. However, the prerequisite to this approach is a sufficiently high number of reports included into the baseline.

$$AD(r) = \frac{|sro \in r|}{\frac{|sdo \in r|(|sdo \in r|-1)}{2}} \tag{22}$$

5.4 Aggregating quality indicators

The aggregation of DQ dimension scores for CTI has to be customizable as described earlier in Sect. 4.2. Adjustable aggregation parameters enable CTI consumers to define the weight of each of the DQ dimensions D in the procedure of calculating a quality indication for each STIX object. For this customizable aggregation, we propose a weighted average (see Eq. 23), where each dimensional score $d_i \in D$ is weighted with a parameter $w_i \in \mathbb{N}$. This parameter w_i can be adjusted by each platform consumer. If no custom value is provided for a dimension d_i, the default weight is $w_i = 1$.

To support consumers' decisions on which report available on a CTI sharing platform to analyze, we additionally propose a report quality indicator calculated following Eq. 24. This score contains the individual DQ object scores $DQ(o)$ and the additional report-level dimension of the appropriate amount of data $AD(r)$. Only the additional DQ dimension's score is weighted in this aggregation with $w \in \mathbb{N}$. The default value for w is again 1. Following this aggregation structure, the weight of each DQ dimension is adjustable by the platform users consuming the respective CTI to ensure that the quality scores represent their individual preference of the dimension's importance.

$$DQ(o) = \frac{\sum_{d \in D} d_i \cdot w_i}{\sum_{d \in D} w_i} \tag{23}$$

$$DQ(r) = \frac{(\sum_{o \in r} DQ(o)) + AD(r) \cdot w}{|r| + w} \tag{24}$$

6 Visualizing quality of CTI

Informing users of CTI about the quality of the intelligence at hand is of crucial importance. This is a vital task in the context of a sharing platform as it allows users to build trust toward the shared CTI. We argue that it is not enough to only inform users about the result of a CTI quality assessment. Instead, the assessment process itself must be transparent for security analysts. Thus, a visual interface should inform them "Why" a report has a specific quality score. As different aspects of CTI quality might also be of varying importance for users, the visual interface could also support parametrization of the quality aggregation as described earlier. Besides the need to inform users about the CTI quality and building trust, their subjective perception of a report's quality is highly relevant for the assessment process. Therefore, a solution is needed to allow them to share their opinion.

Providing a possible path to solve these requirements, we draw upon the idea to make complex threat intelligence exchange formats, like STIX, accessible for human experts through an interactive visual interface. The feasibility and applicability of this approach have been shown in earlier work [29]. In this work, we implemented and evaluated an open-source visual analytics prototype for STIX. We extend this proof of concept by including indicators about the CTI quality in the interface and by implementing functionalities for experts to share their subjective quality assessment where necessary. In the following sections, we briefly introduce the changes made to the original visual interface called Knowledge-Assisted Visual Analytics for STIX (KAVAS). Additionally, we extend the underlying database (CTI Vault) to integrate notions of threat intelligence quality. However, both the database and the visual representation are built to only handle data compliant to the STIX specification. Thus, before including CTI quality into the tool, a solution to represent CTI's quality in the STIX format is needed.

6.1 Integrating quality indicators into STIX

In its current specification, the STIX format has no object types or properties to model indications about the quality of CTI. However, the specification defines the format in a way which allows for the extension of the baseline specification [31]. This opens different possible ways to integrate CTI quality into this format. On the one hand, it is possible to define completely new types of STIX objects. On the other hand, additional properties could be added to the existing SDO and SRO types.

```
{
  ``type'':``x-quality-indicator'',
  ``id'':``x-quality-indicator--1'',
  ``created'':``2019-07-25T09:00:00Z'',
  ``modified'':``2019-07-25T09:00:00Z
     '',
  ``object_ref'':``threat-actor--1'',
  ``measures'': [
    {
      ``dimension'':``Syntactic
         Accuracy'',
      ``type'':``objective'',
      ``score'':0.8
    },
    ...
    {
      ``dimension'':``Reputation'',
      ``type'':``subjective'',
      ``score'':0.7,
      ``rating_count'':14}
    }
  ]
}
```

Listing 2 Exemplary *Quality Indicator* object

Table 1 Definition of the *Measure* custom data type for STIX 2

Property name	Type	Description
dimension (*required*)	String	The dimension for which the measurement is described
type (*required*)	String (*"subjective" or "objective"*)	Describes whether the dimension's score is based on a subjective or an objective metric
score (*required*)	Float	Double-precision number ranging from 0 to 1 describing the current result of the quality assessment for a quality dimension
rating_count (*optional*)	Integer	This property is only needed for "subjective" measures as it describes how many different ratings were given to produce the current score

Table 2 Definition of the *Quality Indicator* custom object for STIX 2

Common properties		
type, id, created_by_ref, created, modified, revoked, labels, external_references, object_marking_refs, granular_markings		
Quality indicator specific properties		
object_ref, measures		
Property Name	Type	Description
type (*required*)	string	The value of this property MUST be "x-quality-indicator"
object_ref (*required*)	identifier	Specifies the STIX Object that is referred to by this quality indicator
measures (*required*)	list of type measure	A list holding all measurements for the different quality dimensions available for the referred-to STIX Object

In any case for each STIX object, the calculated scores for the different quality dimensions need to be documented. To capture the necessary information in a STIX-conformant way, we therefore propose the custom data type *Measure* defined in Table 1. This data type consists of the name of a specific dimension and the object's respective score. It is worth noting that our proposal centers on float values. Nevertheless, scores on an ordinal scale are also possible. Respective conversions can be implemented by defining ranges of float values which refer to a specific ordinal scale (low, medium, and high). Additionally, the custom data type contains the type (subjective or objective) of the dimension. For subjective dimensions, the count of received ratings used to calculate the score can be stored.

We opt to attach a list of measures structured according to the proposed *Measure* data type to a new Custom STIX object. While it is also possible to include this list in any existing STIX object, our proposal aims to maintain a clear separation between actual threat information and the related quality information. Additionally, this proposal produces as less interference as possible with the existing data model. Neither the existing SDOs nor SROs need to be changed. In compliance with the specification, we follow the mechanisms and requirements given to introduce custom objects called *Quality Indicator*. Besides the mandatory *Common Properties*, a number of specific properties are established [31].

Table 2 defines the proposed STIX Custom Object. We include common properties of our *Quality Indicator* object which are mandatory for each SDO. These properties are followed by several specific properties defined for the object. The last part of Table 2 defines allowed data types and values for the specific *Quality Indicator* properties. The *type* attribute must not hold other values than "x-quality-indicator". The *Quality Indicator* object is not connected to any other objects with an explicit SRO but holds a property "object_ref" reflecting the ID of the SDO or SRO for which the object indicates the relevant quality measures. Finally, the object contains a list of "measures" which holds the scores for all the DQ dimensions. The list is formed of the custom *Measure* data type. An exemplary and simplified object is shown in Listing 2.

STIX is an actively maintained CTI standard. Recently, there have been developments that incorporate some aspects similar to our CTI quality concept within the newest STIX2.1 Committee Specification Draft.[5] Most notably, this draft includes an *Opinion* SDO to capture perceptions by CTI consumers about the correctness of a STIX object. The *Opinion* SDO aims to document the level of agreement with the referred-to STIX object(s) on a Likert-type scale ranging from strongly disagree to strongly agree. As can be seen by the purpose and the description of the *Opinion* SDO, this spe-

[5] https://docs.oasis-open.org/cti/stix/v2.1/stix-v2.1.html.

cific STIX object is another prospective option to implement elements of the *Reputation* data quality dimension. Nevertheless, in contrast to our proposed *Quality Indicator* SDO the draft and its *Opinion* SDO fall short to cover a larger CTI quality concept.

6.2 Persisting quality indicators in the CTI Vault

The original database for the CTI visualization is a graph-based approach based on Neo4J.[6] This is quite reasonable as STIX is based in graph-like structure itself. Additionally, the integrity-preserving storage concept proposed by Böhm et al. [29] is most efficiently implemented using this technology. We extend this approach by adding a new database to the architecture. This new database is solely supposed to persist the *Quality Indicator* objects introduced in Sect. 6.1. As described, these objects do not have any explicit connections to other STIX objects via SROs. Their integration would double the number of objects inside the existing database and would certainly affect the performance negatively. Therefore, we decided to avoid storing the quality object inside the existing vault.

Our newly added "Quality Vault" is a document-oriented database (MongoDB[7]) for performance reasons. This additional vault persists the JSON representations of the *Quality Indicator* objects which are directly related to a single SDO or SRO in the CTI Vault via the "object_ref" attribute.

6.3 Displaying quality indicators in KAVAS

Throughout this section, we describe the changes we made to the original visual interface to include visual indications about the quality of STIX artifacts. In Böhm et al. [29], the process of visually analyzing STIX-based CTI with KAVAS starts with a simple drop-down menu to select the report of interest. The drop-down menu contains only the name of the report given by its publisher. This does not disclose any additional information to the analyst whether the report might be of interest or not. We changed this initial view of the KAVAS interface to be more informative and also to give first insight into the quality of the report. The visual interface now contains an expandable list of all available reports from the CTI Vault. The expansion panel for each STIX report consists of three main sections depicted in Fig. 4 informing analysts on the contents and overall quality of a STIX artifact at first glance[8]:

1. At first, a description (if given by the report's producer) gives high-level information on what the report is about.
2. The second section shows which specific STIX objects, both SDOs and SROs, are contained in the report and how often they are present. Object types that are not present in the respective STIX artifact are grayed out. This view fulfills the requirement to provide a view on the homogeneity of a STIX artifact as described in Sect. 5.3 within the quality dimension of *Appropriate amount of data*.
3. The third section gives a very brief and high-level indication on the average quality of the STIX objects and their interconnectedness within the respective report using two gauge displays. This connectedness is represented by the score as described in Sect. 5.3.

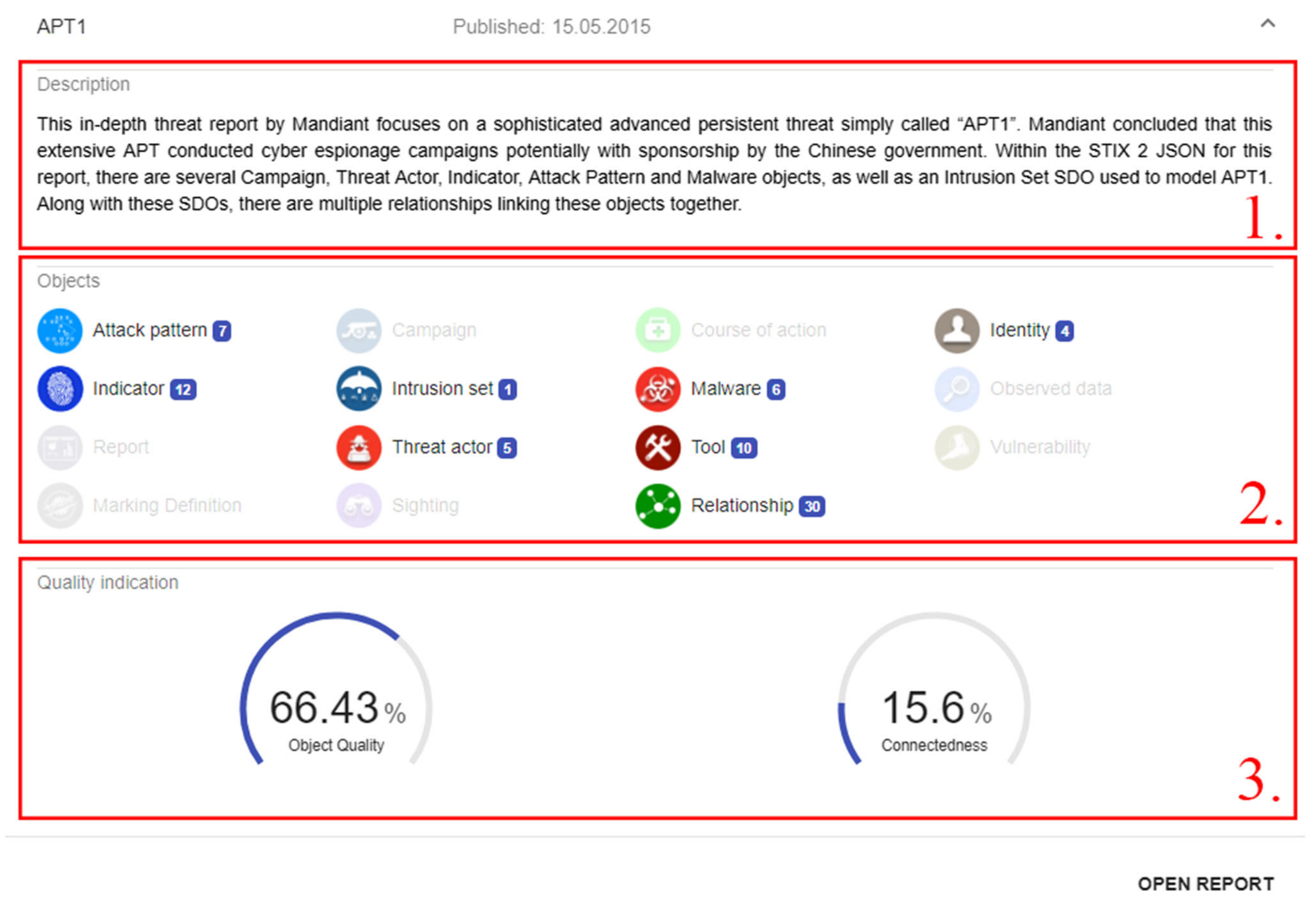
After a STIX report is selected and its graph representation is loaded in the visual interface further changes become apparent, clicking a node or a link of the graph details its information in a details-on-demand card view. The original object card only contained a tab with the attribute values of the selected object and, for SDOs, a tab with its directly linked neighbors in the graph. We now add a quality badge in the header of the object card displaying the aggregated quality score of the dimensions from object and attribute level as described in Sect. 5.4. Furthermore, we add a new tab providing more detailed insight and transparency of the quality measuring. The new object quality tab on the details-on-demand view is shown in Fig. 5. Again, this component is divided into three sections:

1. A gauge visualization of the object's overall quality score aggregated from the scores at the attribute and object level.
2. A section with progress bars indicating the object's score for all described objective dimensions.
3. A third section that holds the indicators for an object's scores of subjective quality dimensions. For this part of the quality tab, we need to both inform the user about the current score and allow them to provide their own subjective quality measurement for the respective dimensions. To do so, we lend from reputation systems and display a rating bar ranging from one to five stars which is a well-known visual metaphor in reputation systems. These rating bars always show the current overall score for the quality dimension (blue stars) in relation to the possible highest rating while also allowing users to click each of the stars to provide their own rating. Numbers in parentheses besides the name of the DQ dimension indicate the count of ratings provided by other users (e.g., the number of subjective assessments on which the current score is based).

[6] https://neo4j.com/.

[7] https://www.mongodb.com/.

[8] Please note that the displayed information is computed based on a test data set which is different from the STIX example in Sect. 3.2.



APT1 Published: 15.05.2015

Description

This in-depth threat report by Mandiant focuses on a sophisticated advanced persistent threat simply called “APT1”. Mandiant concluded that this extensive APT conducted cyber espionage campaigns potentially with sponsorship by the Chinese government. Within the STIX 2 JSON for this report, there are several Campaign, Threat Actor, Indicator, Attack Pattern and Malware objects, as well as an Intrusion Set SDO used to model APT1. Along with these SDOs, there are multiple relationships linking these objects together.

Objects

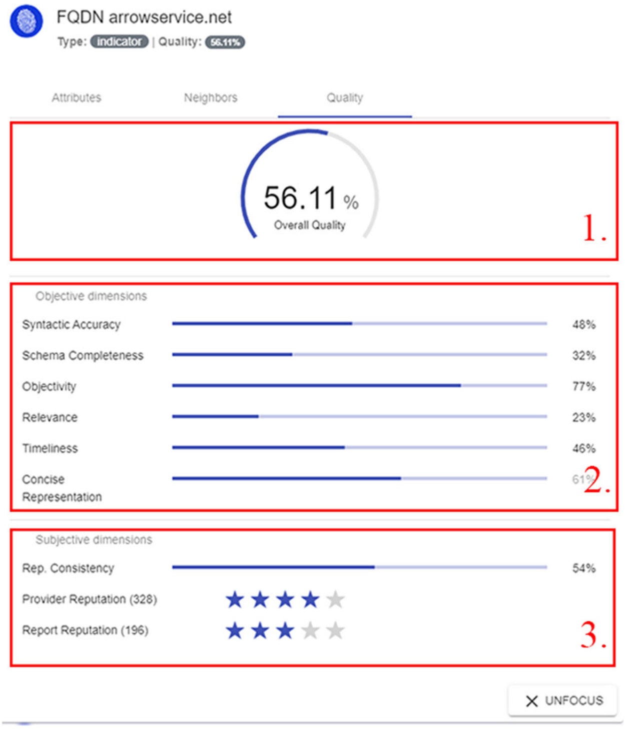
Attack pattern 7
Campaign
Course of action
Identity 4
Indicator 12
Intrusion set 1
Malware 6
Observed data
Report
Threat actor 5
Tool 10
Vulnerability
Marking Definition
Sighting
Relationship 30

Quality indication

66.43 % Object Quality

15.6 % Connectedness

Fig. 4 View of report selection screen



FQDN arrowservice.net
Type: indicator | Quality: 56.11%

Attributes Neighbors Quality

56.11 % Overall Quality

Objective dimensions

Syntactic Accuracy 48%
Schema Completeness 32%
Objectivity 77%
Relevance 23%
Timeliness 46%
Concise Representation 61%

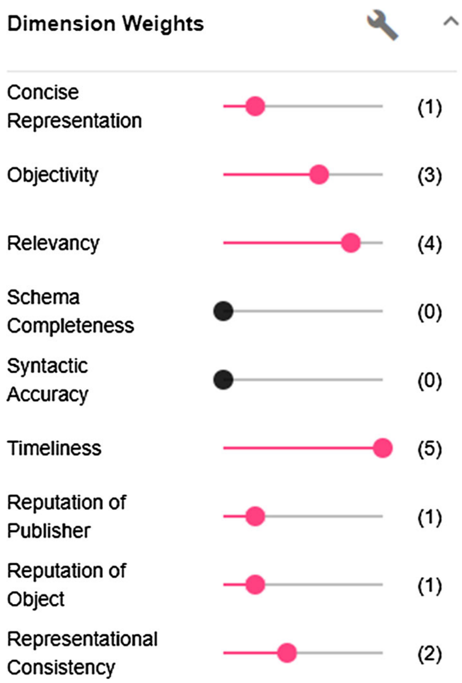
Subjective dimensions

Rep. Consistency 54%
Provider Reputation (328) ★★★★☆
Report Reputation (196) ★★★☆☆

Fig. 5 Quality tab on object card (details-on-demand)

The quality tab fulfills a twofold goal: First, it makes the aggregation of the quality dimensions transparent, and second, it allows collecting user’s input for subjective quality dimensions. We actively decided not to use any color-coding for the scores. Traditionally, respective scores are colored with red (low quality), orange (medium quality), and green (high quality). However, we only aim to inform CTI analysts about the quality scores and do not want to provide any kind of interpretation of low or high score for any quality dimensions. As described earlier, this is mainly because the quality dimensions might be of different interest for different consumers. Therefore, low scores for respective dimensions of an object do not automatically implicate that the object is irrelevant or of low overall quality for the consumer.

In order to allow users to customize the aggregation of quality dimension scores following our previously described bottom-up approach, analysts need a way to define the dimensions’ weights. To provide this functionality, we extend the KAVAS settings dialog with a slider for each quality dimension as depicted in Fig. 6. The default configuration assumes that all dimensions are equally important (e.g., have a weight of 1). Analysts can use the sliders to customize the dimension aggregation according to their preference. If they do not want a specific dimension to have any influence in the aggre-



Dimension Weights

Concise Representation (1)

Objectivity (3)

Relevancy (4)

Schema Completeness (0)

Syntactic Accuracy (0)

Timeliness (5)

Reputation of Publisher (1)

Reputation of Object (1)

Representational Consistency (2)

Fig. 6 Slider for dimension weights

gation, they can assign a weight of 0 and for a dimension with crucial importance, they can accordingly assign a weight of 5. Please note that the metric for dimension aggregation in Eq. 23 does not limit the range for the dimension's weight. However, we chose to limit them in the visual interface to a range from 0 to 5 for more practical feasibility.

6.4 Evaluating the visual display of CTI quality

To validate the visualization approach and to provide first evidence of its suitability, we conduct a number of expert interviews. The main goal of these interviews is to validate that the visual approach helps analysts to understand the DQ of the CTI artifact at hand.

Participants The interviewees are three security experts from different sectors and company sizes. We conduct interviews with two highly experienced security analysts from a big international conglomerate and a medium-sized manufacturing company. The third interviewee is a researcher focusing on CTI sharing formats. Each participant has a medium to high knowledge regarding threat intelligence as all of them deal with information security on a daily basis. None of the participants currently obtains a quality assessment on CTI.

Design and procedure The interviews with the experts are designed following a semi-structured approach and are splitted into the following four phases [32]:

1. *Introduction* Starting the interviews, each participant is questioned for some basic data, their experience, such as knowledge on CTI and DQ aspects. Afterward, each expert is introduced briefly to the STIX format (if necessary) and to the problem of measuring CTI quality. Thereby, the experts are actively asked to criticize any potential issues noticed throughout the following interview phases.
2. *Measuring CTI quality* In this phase, we aim to get additional feedback on the individual dimensions and the configured metrics for quality assessment of STIX artifacts (Sects. 4 and 5). Although the dimensions and the metrics are already the result of an iterative process where we collaborated with researchers and practitioners, an additional evaluation of these results is performed in this phase. The selected dimensions, their structure, and the configured metrics are discussed with the interviewees to identify whether they support the relevance of the proposed DQ measurement approach. We also ask the participants what aspects of the dimensions and metrics might need a more detailed explanation and whether they think that the metrics are comprehensible for security analysts without much prior knowledge in the DQ area.
3. *Visualizing CTI quality* The focus of this phase is to test the suitability of the proposed visualization approach. To enable the interviewees to work with the DQ visualization, we make use of sample STIX reports provided by the OASIS consortium. Prior to the interviews, these reports were manually fed into the existing KAVAS tool and enriched with the DQ measures. During the interviews, the participants can access the STIX reports through the extended KAVAS tool as described in Sect. 6. The main goal in this interview phase is to identify whether the proposed visualization elements to display the CTI quality are actually helpful for security analysts. We ask the interviewees whether the proposed DQ metrics are comprehensible with the chosen visualization elements and what further aspects they think would enhance the understanding of DQ assessment within CTI.
4. *Wrap-Up* The last phase of the interviews is dedicated to a summarizing discussion. Here, we discuss with the participants whether an implementation of the proposed metrics and the respective visualization approach would be applicable to operative deployment and the conditions thereto. Finally, we collect a list of ideas and features the interviewees find useful for improving our approach.

Results The interviews lasted between 45 to 75 minutes. The results of the conducted interviews are presented in the following, divided according to the four interview phases described before:

Table 3 General information on the interview participants

	Position	Business branch	Organization's size	CTI knowledge	DQ knowledge
#1	Senior security analyst	Manufacturing	ca. 400.000	High	Medium
#2	Head of security information management	Manufacturing	ca. 15.000	Low	Medium
#3	Security researcher	Academia	ca. 5.000	High	Medium

1. *Introduction* The results of the introduction phase are summarized in Table 3 giving an overview on general information about the interviewees.
2. *Measuring CTI quality* Above all, the interviewees unanimously stress the importance of metrics for quality within the field of CTI. Valuable and actionable CTI is stated to be highly dependent on quality and currently more often than not CTI is of low quality. A recurring theme mentioned in this phase by multiple interviewees is the interpretation of CTI quality. It is pointed out that the implementation of metrics for CTI quality by sharing platforms would benefit significantly from indication of low- and high- quality reference scores. Another identified theme is usability of DQ dimensions and metrics for CTI. Here, formally sound metrics, the chosen naming convention of DQ dimensions based on existing academic work and security analysts without DQ or mathematical background, stand opposite each other. Comprehensive explanations are seen as one approach to foster security analysts' understanding of the precise meaning of CTI quality dimensions and metrics.
3. *Visualizing CTI quality* All interviewees agree on the necessity to provide easy access to CTI quality through the use of visualization elements and validate our visualization approach. All interviewees agreed that the chosen visual representation allows for a quick recognition of CTI quality. They also uniformly considered the possibility to include subjective perceptions with means similar to reputation systems very helpful. Nevertheless, the interviewees name different extensions to the current visualization. For one, in-depth information about the DQ dimensions, the metrics, and possible interpretation is highlighted. Additionally, the showcased visualization includes percentages numbers and numeric weighting factors which could instead be visualized on a Likert-type scale. Another proposed extension targets the causal nature of low-quality scores. Visualization elements to detect improvements and eventually improve the CTI quality further are perceived as helpful. As one interviewee points out, user groups (e.g., system administrator or standard user) could be defined, given different permissions and thus see different visualizations.
4. *Wrap-up* In the final phase, the interviewees often come back to the timeliness dimension. The proposed metrics for this DQ dimension needed additional explanations with regard to STIX specifics (i.e., Sighting SDO). Ideas and features mentioned by the interviewees to extend our work cover guidance to improve CTI quality and quality filtering with visualization elements. For instance, visual recommendations to reach a higher CTI quality (with or without prior knowledge about quality details) might be added to the current reactive assessment.

Overall, the interviewees' feedback indicates the valuable contribution of measuring and visualizing CTI quality. In particular, the dual approach itself (measure and visualize) is assumed to reduce complexity, lower quality assessment barriers, and foster CTI utilization. With regard to the implementation within a CTI sharing platform, we draw the conclusions that 1) there needs to be discussion on usability and adequate naming of DQ dimensions, 2) reference values are crucial for CTI quality interpretation, and 3) visual elements and textual explanations must be combined to avoid ambiguity.

7 Conclusion and future work

This work shed light on the assessment of DQ dimensions in the context of CTI. Nonetheless, there are further areas where research needs to be intensified and extended to.

7.1 Conclusion

Recent developments in the cyber threat landscape urge organizations to join forces against the adversaries. Collaboration based on the exchange of available threat intelligence arises as one of their most effective weapons. CTI sharing leveraged by respective platforms helps to spread knowledge about current threats. However, respective formats are oftentimes complex and large leading to a lack of readability for domain experts. Therefore, it is a vital task to help experts understand the CTI, for example, by providing visual representations. CTI can only be effective when security experts are able to comprehend it quickly and efficiently. Another issue hindering the effectiveness of CTI is the missing quality control on sharing platforms. This lack of DQ management mostly stems from missing proposals to measure CTI quality in the first place.

Our studies cumulated within this work constitute a necessary first step into this direction. This includes the two focal points of measurement and visualization of threat intelligence quality. Existing academic work proposed sets of possibly relevant quality dimensions as well as high-level requirements for CTI quality assessment. Although calling for an inclusion of quality assessment and assurance into the world of CTI sharing, up to now there are no proposals for actual quality metrics applicable to CTI. Therefore, proposing a relevant set of quality dimensions and configuring respective metrics for a specific CTI format is a necessary step toward actionable CTI quality assessment. The proposed dimensions and metrics can help to build a cohesive quality management methodology for CTI based on the STIX data format. Most of our findings regarding suitable as well as not applicable DQ dimensions or metrics can also be applied to other CTI formats. It is possible to think of additional, more specific dimensions which could be defined to assess quality of threat intelligence. However, in this work we define a base set of dimensions that originate from existing and widely agreed-upon DQ dimensions. This base set can easily be extended, and detailed metrics can complement our proposed ones if necessary.

Besides the definition of metrics to measure CTI's quality for relevant dimensions, we also showed how this quality assessment can be made transparent to users of a sharing platform. Transparency herein supports both building trust for the available information and making informed decisions about which CTI artifact is worth analyzing. This is important as current sharing platforms already hold an unmanageable amount of threat intelligence. Informing potential consumers of an artifact about its quality is a helpful decision support for the consumer. The visual display of an object's overall quality including the respective scores for individual quality dimensions helps consumers to understand how the DQ measurement result was reached. Additionally, it provides a way to collect important input from users for subjective quality dimensions. We therefore also show how human CTI analysts can be included into the quality assessment.

7.2 Future work

Our work can be seen as a first step into the direction of measuring CTI quality. However, we can identify several topics demanding additional research effort.

We are among the first to propose a cohesive set of applicable CTI quality dimensions. Therefore, these dimensions might be subjected to changes as more knowledge is gained about CTI sharing processes, platforms, and associated stakeholders. One dimension which needs further attention is the *Appropriate amount of data*. The proposed metric is a first approach toward a highly complex issue. It is difficult to define which amount of data—either data regarding STIX objects or the information described by these objects—is appropriate. Thus, we propose a simple metric to give domain experts an indication of the data contained in a STIX report. The DQ metric for the appropriate amount of data should be further detailed upon analysis and verification with CTI platform data. Furthermore, the metrics to evaluate quality should be reconfigured for other CTI formats and integrated into a cohesive data quality management methodology for CTI.

After formally configuring the metrics for the selected quality dimensions, those metrics should be implemented into an actual sharing platform. Up to now, we only tested them in a small scaled environment. A complete implementation will likely raise further issues about the selection of suitable algorithms and the control of user participation and intentions which go beyond the core DQ assessment and have not been addressed in this work. Warranted through an implementation, the extension of some proposed dimensions can become feasible as more information about the requirements will be available. Implementing and extending the dimensions and metrics are necessary steps to finally build a cohesive methodology for quality assessment of CTI including processes to assure and improve quality of artifacts on a sharing platform.

Acknowledgements Open Access funding provided by Projekt DEAL.

Funding Part of this research was supported by the Federal Ministry of Education and Research, Germany, as part of the BMBF DINGfest project (https://dingfest.ur.de). Part of this project has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 830927.

Compliance with ethical standards

Conflict of interest All authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Symantec Corporation.: Internet security threat report 2019 (2019). https://www.symantec.com/content/dam/symantec/docs/reports/istr-24-2019-en.pdf
2. Riesco, R., Villagrá, V.A.: Leveraging cyber threat intelligence fora dynamic risk framework. Int. J. Inf. Secur. **18**, 715–739 (2019)
3. Ponemon Institute LLC.: Live threat intelligence impact report 2013 (2013). https://www.ponemon.org/blog/live-threat-intelligence-impact-report-2013-1
4. Ring, T.: Threat intelligence: Why people don't share. Comput. Fraud Secur. **2014**(3), 5 (2014)
5. Sillaber, C., Sauerwein, C., Mussmann, A., Breu, R.: Data quality challenges and future research directions in threat intelligence sharing practice. In: Proceedings of the 2016 ACM on Workshop on Information Sharing and Collaborative Security - WISCS'16, pp. 65–70. ACM, New York (2016)
6. Sillaber, C., Sauerwein, C., Mussmann, A., Breu, R.: Towards a maturity model for inter-organizational cyber threat intelligence sharing: A case study of stakeholder's expectations and willingness to share. In: Proceedings of Multikonferenz Wirtschaftsinformatik (MKWI 2018), pp. 6–9. Springer, Heidelberg (2018)
7. Tounsi, W., Rais, H.: A survey on technical threat intelligence in the age of sophisticated cyber attacks. Comput. Secur. **72**, 212–233 (2018)
8. Juran, J.M., Gryna, F.M.: Juran's Quality Control Handbook, 4th edn. McGraw-Hill, New York (1988)
9. Jøsang, A., Ismail, R., Boyd, C.: A survey of trust and reputation systems for online service provision. Decis. Support Syst. **43**(2), 618 (2007)
10. Dandurand, L., Serrano, O.S.: Towards improved cyber security information sharing. In: 2013 5th International Conference on Cyber Conflict (CYCON 2013). IEEE Computer Society Press, Los Alamitos (2013)
11. 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 - WISCS '14, pp. 61–69. ACM, New York (2014)
12. Kokulu, F.B. Soneji, A. Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., Ahn G.J.: Matched and mismatched socs: a qualitative study on security operations center issues. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security (Association for Computing Machinery, New York, NY, USA, 2019), CCS '19, pp. 1955–1970. https://doi.org/10.1145/3319535.3354239
13. Sauerwein, C., Sillaber, C., Mussmann, A., Breu, R.: Threat intelligence sharing platforms: an exploratory study of software vendors and research perspectives. In: Proceedings of the 13th International Conference on Wirtschaftsinformatik, pp. 837–851. Springer, Heidelberg (2017)
14. Menges, F., Pernul, G.: A comparative analysis of incident reporting formats. Comput. Secur. **73**, 87–101 (2018)
15. Piazza, R., Wunder, J., Jordan, B.: StixTM version 2.0. part 2: Stix objects (2017). https://docs.oasis-open.org/cti/stix/v2.0/stix-v2.0-part2-stix-objects.html
16. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. **41**(3), 1 (2009)
17. Skopik, F., Settanni, G., Fiedler, R.: A problem shared is a problem halved: a survey on the dimensions of collective cyber defense through security information sharing. Computers & Security **60**, 154–176 (2016)
18. Batini, C., Scannapieco, M.: Data and Information Quality: Dimensions, Principles and Techniques. Springer, Cham (2016)
19. Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Commun. ACM **39**(11), 86 (1996)
20. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5 (1996)
21. Redman, T.C.: Data Quality for the Information Age. Artech House Publishers, Norwood (1996)
22. Umbrich, J., Neumaier, S., Polleres, A.: Quality assessment and evolution of open data portals. In: 2015 3rd International Conference on Future Internet of Things and Cloud (FiCloud), pp. 404–411. IEEE Computer Society Press, Los Alamitos (2015)
23. Wang, R.Y., Storey, V.C., Firth, C.P.: A framework for analysis of data quality research. IEEE Trans. Knowl. Data Eng. **7**(4), 623 (1995)
24. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Commun. ACM **45**(4), 211 (2002)
25. Batini, C., Palmonari, M., Viscusi, G.: The many faces of information and their impact on information quality. In: Proceedings of the 17th International Conference in Information Quality (ICIQ 2012), pp. 212–228. MIT, Cambridge (2012)
26. Sänger, J., Richthammer, C., Pernul, G.: Reusable components for online reputation systems. J. Trust Manag. **2**(5), 1 (2015)
27. Gascon, H., Grobauer, B., Schreck, T., Rist, L., Arp, D., Rieck, K.: Mining attributed graphs for threat intelligence. In: Proceedings of the 7th ACM on Conference on Data and Application Security and Privacy, pp. 15–22. ACM, New York (2017)
28. Chaturvedi, I., Cambria, E., Welsch, R.E., Herrera, F.: Distinguishing between facts and opinions for sentiment analysis: survey and challenges. Inf. Fus. **44**, 65 (2018)
29. Böhm, F., Menges, F., Pernul, G.: Graph-based visual analytics for cyber threat intelligence. Cybersecurity (Cybersecurity) **1**, 1 (2018)
30. Heinrich, B. Kaiser, M. Klier, M.: How to measure data quality? A metric-based approach. In: ICIS 2007 Proceedings pp. 108–122 (2007)
31. Piazza, R., Wunder, J., Jordan, B.: StixTM version 2.0. part 1: Stix core concepts (2017). https://docs.oasis-open.org/cti/stix/v2.0/stix-v2.0-part1-stix-core.html
32. Lazar, J., Feng, J.H., Hochheiser, H.: Research Methods in Human–Computer Interaction. Morgan Kaufmann, Burlington (2010)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.