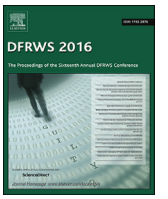
Digital Investigation 18 (2016) S125–S137

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin



DFRWS USA 2016 — Proceedings of the 16th Annual USA Digital Forensics Research Conference

CuFA: A more formal definition for digital forensic artifacts

Vikram S. Harichandran, Daniel Walnycky, Ibrahim Baggili*, Frank Breitinger

Cyber Forensics Research and Education Group (UNHcFREG), Tagliatela College of Engineering University of New Haven, West Haven, CT, 06511, United States

Keywords:
Forensic artifact
Digital forensics
CybOX
Curated forensic artifact
CuFA
Artifact definition
Survey
Cyber forensics
Taxonomy
Ontology

A B S T R A C T

The term “artifact” currently does not have a formal definition within the domain of cyber/digital forensics, resulting in a lack of standardized reporting, linguistic understanding between professionals, and efficiency. In this paper we propose a new definition based on a survey we conducted, literature usage, prior definitions of the word itself, and similarities with archival science. This definition includes required fields that all artifacts must have and encompasses the notion of curation. Thus, we propose using a new term – curated forensic artifact (CuFA) – to address items which have been cleared for entry into a CuFA database (one implementation, the Artifact Genome Project, abbreviated as AGP, is under development and briefly outlined). An ontological model encapsulates these required fields while utilizing a lower-level taxonomic schema. We use the Cyber Observable eXpression (CybOX) project due to its rising popularity and rigorous classifications of forensic objects. Additionally, we suggest some improvements on its integration into our model and identify higher-level location categories to illustrate tracing an object from creation through investigative leads. Finally, a step-wise procedure for researching and logging CuFAs is devised to accompany the model.

Introduction

Currently, the use of the term “artifact,” or “artefact” (United Kingdom spelling), in relation to digital information and cyber/digital forensics embodies a variety of meanings depending on the context used as well as the perspective of the user. The term has generally been adopted within the cyber forensics domain for items of interest that help an investigation move forward. Notwithstanding, the lack of a formal definition and sound ontology is holding the field back from forming standards to keep pace with cybercrime (Brinson et al., 2006). Note that the term should not be confused with the software development interpretation of the word (a tangible by-product produced during software development, especially pertaining to such methods/processes).[1]

Without a systematic ontology, scientists and practitioners have different ideas of how knowledge is related within the context of their situations. Ontology provides an essential “unifying map of concepts and relationships” (for more explanation on the importance and creation of ontologies/taxonomies see Malafsky and Newman (2009)). Chiefly, professionals (in different subdomains) cannot easily share evidence and often are forced to rely exclusively on their own past experience during investigations,

* Corresponding author.
E-mail addresses: vhari2@newhaven.edu (V.S. Harichandran), dwaln1@unh.newhaven.edu (D. Walnycky), ibaggili@newhaven.edu (I. Baggili), fbreitinger@newhaven.edu (F. Breitinger).
URL: http://www.unhcfreg.com/, http://www.FBreitinger.de/

[1] http://forensicswiki.org/wiki/Computer_forensics#Artifact (last accessed Feb. 2, 2016).

http://dx.doi.org/10.1016/j.diin.2016.04.005

which may cause missed evidence or leads. This becomes extremely important with the ubiquity of devices and software applications used today. Adopting an ontological system should increase the ability to bring to light connections investigators are unaware of, such as linked cases that involve the same criminal, or missing data in a specific location that indicates system tampering.

In addition to the ontology, it is important to develop a standardized taxonomy so that reports can be developed by software/investigators easily via a procedure for the process of researching and handling items (curation). By using dynamic (optional) and required fields, artifacts extracted using various tools would be directly comparable. Currently this is not the case for cyber forensics (e.g. "SerialNum" on Windows vs. "Serial Number" on OSX), even though such classification exists for biological forensics (Brady et al., 2014). The open-source CybOX project[2] is one increasingly popular attempt at standardizing such fields. Object types encapsulate these fields making items placed in them close to mutually exclusive, but like the prior example there lacks details that help experts enter data on cyber items (files, processes etc.). Conventions are especially lacking with respect to presentation of evidence in courts (Bariki et al., 2011).

Our contribution aimed to solve the aforementioned challenges (standardization/cohesive understanding and standardization of practitioner-oriented information exchanging). Primarily, a survey was designed to ask practitioners and researchers how they would define a "digital forensic artifact". Based on this, previous adoptions in academic literature, and the domain of archival science we accomplished the following:

1. Proposed a more concrete, unified linguistic definition, assigning it a new name: Curated (digital) Forensic Artifact (CuFA).
2. Using survey responses and our proposed definition, we designed an ontological model for curation of artifacts that involves a procedure and sets the requirements for an object to be considered a CuFA.
3. Presented a manner for implementing the higher-level ontology in conjunction with a low-level schema (CybOX) resulting in a searchable database organized by dynamic, taxanomic fields and tags/flags.

The structure of this paper is as follows: first, we cover past definitions/usage of the term "artifact" (Developing a definition section), review previous ontologies (Outlining an ontological model section), and deliver a brief comparison to archival science. Next, our survey methodology (Methodology section) and design (Survey design section) are introduced, followed by the data in the Results section. Using these findings we propose a definition and an ontological model based on this definition in the Proposed definition and model section. Finally, discussion and suggestions for future work are presented to the reader.

[2] http://cyboxproject.github.io (last accessed Feb. 2, 2016).

Previous work

Developing a definition

This section reviews past definitions of the term "artifact" in the context of digital evidence, the types of items both researchers and organizations have used the term to describe (including the perspective that drives these usages), and previously proposed ontologies.

Definitions

All definitions listed below are word-for-word. Merriam-Webster Dictionary (2015) defines "artifact" as:

- An accidental effect that causes incorrect results.
- Something characteristic of or resulting from a particular human institution, period, trend, or individual.
- A product of artificial character (as in a scientific test) due usually to extraneous (as human) agency.

Oxford Dictionaries (2015) lists:

- An object made by a human being, typically an item of cultural or historical interest.
- Something observed in a scientific investigation or experiment that is not naturally present but occurs as a result of the preparative or investigative procedure.

Dictionary.com (2015)'s definitions include:

- Any object made by human beings, especially with a view to subsequent use.
- A substance or structure not naturally present in the matter being observed but formed by artificial means.
- A spurious observation or result arising from preparatory investigative procedures.
- Any feature that is not naturally present but is a product of an extrinsic agent, method, or the like.

More specific definitions were obtained from the CybOX project (MITRE Corporation, 2015):

- An object produced or shaped by human craft, especially a tool, weapon, or ornament of archaeological or historical interest.
- A phenomenon or feature not originally present or expected and caused by an interfering external agent, action, or process.

Another digital-scoped definition came from the Scientific Working Groups on Digital Evidence and Imaging Technology (SWGDE/SWGIT, 2015):

- Information or data created as a result of the use of an electronic device that shows past activity.

There were a few instances where papers made explicit attempts to bound their usage of the term, and therefore provided a definition. In one case it was stated that artifacts should not be confused with Indicators Of Compromise

(IOCs), items that signify a system's compromise, as their intent is different and they represent pure data without logic, i.e. a system state rather than malware state (Castle, 2014b). The example Castle gave: an IOC might be an executable that contains a string "evil" or is signed "stolen cert," while an artifact could be the location where user runkeys are located (HKEY_USERS\%%users.sid%%\Software\Microsoft\Windows\CurrentVersion\Run*Ω). Wikipedia contradicted this by using the word "artifact" in the definition of IOC.[3] Castle's IOC definition also disagrees with the previous usages presented in the Perspectives and usage section. SysAdmin, Audit, Networking, and Security (SANS) defines an artifact as a "combination of description, location, and interpretation" (Castle, 2014a).

The commonality between these definitions appears to be observed artificiality/external force, antecedent temporal relation, and exceptionality (based on either accidental procurement, rarity, or a person's interest). Including the word "forensic" at the beginning of the term adds legality and science to this list. These cannot be used exclusively to form a definition, however – academic and community usage must be examined.

Perspectives and usage

Citations of the term "artifact" in cyber forensics have varied based on the professional goals of the users' subdomains and the tasks they were performing. Reviewed papers (see the table in the Appendix for the full list of papers and perspectives) used the word mostly in an ad hoc manner that reflected the concept of exceptionality via personal interest; thus, this perspective was the most variant and the term took a different specific meaning in each paper (e.g. log data in Yasin and Abulaish (2013) vs. installation/runtime/deletion behaviors in Lim et al. (2010)).

A second trend recorded was that of investigators. Usages of the term in these papers emphasized looking generally for "what you want to know" in order to further an investigation and, consequently, had a broader intention than the academic standpoint. Between these two extremes laid the perspective of those who design, manufacture, and test tools. The primary motive behind this view is the objective of standardizing objects for tagging (or filling in fields/checkboxes), reporting, comparison (exporting to share), and increased investigative efficiency. Note that although this seems similar to the investigative stance, these papers detracted the logical, conceptual aspects described above and attempted to focus on the location and data itself.

Table 1 shows an excerpt from the table in the Appendix. Each paper's focus was categorized into one of the above perspectives, or the collection perspective described in the next section, so that the various mindsets could be weighted in our proposed model. The investigative ethos was usually used in conjunction with another one and therefore cannot be found in the table; we declare the more prevalent view. Note that papers marked with an asterisk did not have explicit categorization of items and required the authors to devise educated groupings.

Table 1
Excerpt from the Appendix to exemplify the structure of the full table.

Items	Category	Paper & perspective
Apple system log; Crash reporter; Diagnostic messages; FSEvents API; Preference settings; Saved application state; Spotlight; Swap files/paging/cache; Temporary data;	OSX	Sandvik, 2013* Researcher
Prefetch; Thumbnail cache; Paging file; Registry; Windows search;	Windows	
Bash history; GVFS virtual file system; Recently used; X session manager;	Linux	

Note that papers marked with an asterisk did not have explicit categorization of items and required the authors to devise educated groupings.

Outlining an ontological model

Casey et al. (2015) reviewed past ontologies and stressed that "querying data on the basis of high-level behaviors […] can be more powerful than just searching for low-level digital artifacts". In this section, we highlight the advantages and disadvantages of these ontologies and introduce the CybOX model. These will be used to feature a basic ontological model, delivered in the Proposed definition and model section, stemming from our proposed definition.

Traditionally, "ontology" involves the study of existence, the categories of being, and their relationships. However, in computer science the "intention is distinct: to create engineering models of reality, artifacts which can be used by software, and perhaps directly interpreted and reasoned over by special software called inference engines, to imbue software with human level semantics" (Poli et al., 2010). This form of ontology is sometimes referred to as "Little o" ontology, while "Big O" ontology signifies the philosophy-centered definition. There is some overlap between the two, but we were primarily concerned with "Little o". Note, hereafter we use the term "model" as an umbrella term for both an ontology (we define this as high-level) and a taxonomic schema (we define this as technical/low-level).

Ontologies

Before reviewing the conceptual ontologies and technical schemas proposed in the last few years it is important to understand what problems these models attempt to solve. Knowledge and correlation are the main concerns for investigators, i.e. knowing where artifacts are located, knowing how they can assist a case, and connecting artifacts across locations/devices when they may be recorded in different ways (Brady et al., 2014). As stated in the Introduction, increasing the chances of finding leads an investigator is unaware of is also a top priority, whether this be an unfamiliar file type, missing data that indicates a system's state has been changed, or a criminal's modus operandi. A feature many of the following ontologies have is extensibility – the advantage of representing low-level

[3] https://en.wikipedia.org/w/index.php?title=Indicator_of_compromise&oldid=666037196 (last accessed Feb. 2, 2016).

taxonomic data in a way that can be utilized flexibly by high-level ontologies. We categorize these models as a fourth type of perspective in addition to the perspectives presented in the Developing a definition section, that of a (database) collector/designer.

The Forensic Wiki[4] represents a loose catalog of tools, types of digital objects obtained from them, and general cyber forensics topics. However, Brady et al. (2014) identified that its "value would be further enhanced if it used some form of classification or tagging system that allowed examiners to readily access what artifacts were available and how these could be linked across its various categories". The authors suggested to solve these issues by proposing the Digital Evidence Semantic Ontology (DESO), an investigative perspective of data that views artifacts through the scope of location or type superclasses.

In DESO, superclasses inherit other subclasses and attributes. For example, the location class may inherit devices, file systems, and operating systems subclasses (describing specific categories of locations), while the type class may inherit device identifier and logical identifier subclasses. When attempting to further an investigation the two primary classes accompany each other. If an investigator has obtained a specific artifact they may use the location class to look for that artifact type across different devices, file systems, and operating systems; alternatively if the type class is the same for artifacts extracted from various locations they can be compared directly.

Other high-level ontologies include Structured Threat Information eXpression (STIX), Digital Forensics Analysis eXpression (DFAX), and Unified Cyber Ontology (UCO) (Casey et al., 2015). STIX uses CybOX (presented in the next subsection) to represent technical details (e.g. malicious IPs) in a manner that mirrors subdomain-specific information such as threat actors. It is the predecessor to DFAX which also attempts to use a third-party schema within a broader ontology to capture more procedural aspects such as chain-of-custody, case management, or processing. Fields such as *ActionPattern* and *ActionLifecycle* allow users to adopt them for documenting the investigative process, and fields for recording event times allows piecing together timelines and collusion between criminal entities. UCO is an ontology illustrating even more abstract concepts that are linked across the cyber forensics domain. It requires the usage of a lower-level schema and potentially could use more than one schema at the same time for different subdomains.

Schemas

In this subsection we briefly describe the pros/cons of the following schemas: XML Information Retrieval Approach to digital Forensics (XIRAF), Digital Forensics XML (DFXML), and Cyber Observable eXpression (CybOX). XIRAF was a prototype for an XML-based schema proposed by the Netherlands Forensic Institute, but its use of parent–child relationships between objects limited its flexibility and it did not gain anticipated prominence outside of the Netherlands. A subsequent XML-based proposition was DFXML, also an attempt to introduce a structure to the presentation of objects (Garfinkel, 2012). Although the format enabled cross-platform comparison and sharing, it still has not been adopted as a standard, perhaps because XML is a verbose language or because some think it is too limiting (without an accompanying ontology).

Recently, CybOX has gained popularity due to its open-source nature and rigorous classification scheme for objects. CybOX utilizes a long list of required and optional attributes for each object type, which it classifies mainly by where the artifact came from conceptually. Each object is given a Globally Unique Identifier (GUID) to make it easily searchable in a database. One concern CybOX addresses is recording the state of a system before and after an event (e.g. version of a file). Differences can be logged specifically (new file created, timestamp) or statistically (similarity digest). CybOx has begun to be implemented in Trusted Automated eXchange of Indicator Information (TAXII) and other models (Casey et al., 2015). Nevertheless, it lacks the ontological vantage point of high-level ontologies like DFAX – thus, we use it as the low-level building block for our model.

Archival science

Novel work by Dietrich and Adelstein (2015) made comparisons between the fields of cyber forensics and archival science. Both fields use procedures involving acquiring, authenticating, and preserving items in a way that minimizes alterations (and documents them). Aside from maintaining their integrity, items must be able to be easily retrieved for future examination and analysis. As stated by the authors, this is one area where cyber forensics differs from archival science: "most criminal forensic organizations have no long-term data preservation and maintenance policy beyond physical storage". Thus, this should also be considered when attempting to develop a definition; the aspect of curation is what gives items the name "artifact" and sets them apart from items not analyzed within procedures followed by experts.

Survey

Methodology

The following basic methodology was applied in carrying out the survey:

1. Performed comprehensive literature review which informed the researchers that there was neither a consensus in the usage of the term "artifact" nor a concrete definition.
2. Designed a survey around asking respondents to define the term and list possible categories/fields that would help organize such items.
3. Obtained a category two exemption from the Institutional Review Board (IRB) at the University of New Haven restricting the survey from recording participant

[4] http://forensicswiki.org (last accessed Feb. 2, 2016).

identification information or behavior, and disclaiming that it posed risk or harm to subjects not encountered in everyday life.
4. Distributed the survey via list servers and LinkedIn.
5. Retrieved data by exporting the coded responses to an XLSX file from the survey system.
6. Analyzed the data for creation of the definition and ontology in conjunction with past work.

Survey design

Questions were formulated based on what the authors found in literature – a missing cohesive definition for the word "artifact" in the context of the domain, and absence of a comprehensive ontology for entering and organizing such items based on static and dynamic fields. There were two iterations of the survey and a testing round before opening it to the community. The survey consisted of 70 questions:

- 54 Likert scale
- 12 free response
- 4 multiple choice

According to IRB regulations at our institution, participants could not be forced to answer any single question. The target audience was all professionals in the field who had encountered items referred to as "artifacts".

Results

A brief note about the Likert scale figures in this section: Each bar represents one Likert scale question; approximate percentages for each answer selection are displayed within their respective segment; and the number of respondents per question is visible in brackets to the right of each bar.

The survey was open for 2 months before data was exported from the survey system.[5] There were 87 respondents, but 50 of them were excluded for only answering demographic questions; the results summarized below only account for the other 37 participants. It is likely this occurred because these participants wanted to just view the survey questions. A power calculation was not performed for the data because it is mostly descriptive/qualitative, lacking any statistical inferences; we recommend the results be used for determining the effect size for any followup work. As seen in Table 2, more than half of the respondents were Americans, had at least 7 years of experience in digital forensics, and were over the age of 34. Respondent expertise can be viewed in Fig. 1.

Since there was no strong agreement on the definition of a forensic artifact, the responses in the survey spanned a myriad of positions. The general themes are shown below, many of which involve direct quotations. In the following results parenthetical numbers next to responses indicate the number of times a general idea or specific words were mentioned.

[5] Raw data, tabulation, graphs, and the survey itself are publicly available on our website: http://www.unhcfreg.com.

Table 2
Numbers in the table are rounded and thus may exhibit rounding error. The following was not disclosed: two people did not rate their expertise in computer forensics, one did not rate expertise for non-traditional forensics, two didn't describe their experience in the field, and one person did not disclose their gender. These percentages only account for the 37 that answered the non-demographic questions.

	Percentage
Age	
18–24	3
25–34	30
35–44	24
45–54	30
55–64	11
65–74	3
75 or older	0
Gender	
Female	16
Male	81
Other	3
Country	
Antigua and Barbuda	3
Argentina	3
Canada	5
Germany	8
India	5
Russia	3
Togo	3
Turkey	3
United Arab Emirates	3
United Kingdom	14
United States	51
Region	
North America	57
Europe	22
Asia	8
Middle East	5
Africa	3
Caribbean	3
South America	3
Years work experience in digital forensics	
1–3 years	24
4–6 years	22
7–9 years	16
10 years or more	38

- Something that has "evidentiary value" in a legal proceeding (7).
- The results of "applying digital forensic (analysis) techniques" (4).
- Byte stream/sequence (2).

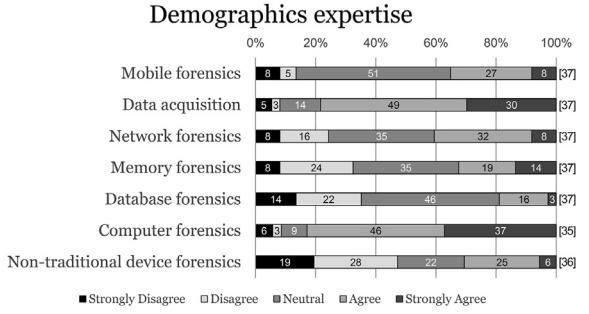


Fig. 1. Results of demographics questions which asked respondents how much they considered themselves experts in the stated subdomain.

- Something of probative interest/yielding information about a digital device (2).
- Digital item/data (2).
- Smallest unit of evidence that can make sense for a digital investigation (timestamp, database entry, etc) (2).
- Something used to reconstruct a crime/events (1).
- File states (1).
- An extraction with an established “data type” (1).
- Semantically annotated metadata (1).

As mentioned in the Introduction, it was important to find a procedure to research and process items. Two separate free response questions asked respondents to reveal their “investigative process” with familiar and unfamiliar artifacts. Four procedures were mentioned for the first question (familiar) and six procedures for the second (unfamiliar). Since mostly similar steps were stated in both questions, we combined these into a proposed procedure for general guidance, which we hope will serve as a method of standardization to be taught to training professionals:

1. Acquire (identify which tool the artifact came from).
2. Backup.
3. Check database to see if encountered before (this can be done by comparing hashes or fields).
4. If familiar, do quick search in artifact database to see if methods used previously are still applicable/effective. If they are, use them, then jump to step 8. If they aren't or the artifact is unfamiliar continue to next step.
5. Classify into a category using the proposed ontological model and catalog/extract artifact qualities (taxanomic fields used in schemas).
6. Attempt to use techniques effective for that category. If ineffective, repeat steps 4–6 until effective and skip to step 8.
7. If no effective techniques are encountered try reconstruction to see if the artifact can be recreated or reverse engineered. Usage of a hex editor may be useful.
8. After a technique is successful in analyzing, repairing, isolating, or rendering the artifact harmless the process should be documented (with all relevant artifact fields) and outputted to a report.
9. Examine the system for associated artifacts based on what was discovered/learned. This may involve searching the database for artifacts of the same type, or using the pointers in the artifact's database entry to browse potentially related artifacts in other locations.
10. Prepare the reports of each (type of) artifact for supporting a legal case.

Finally, the survey also aimed to devise a schema for organizing and archiving items through the identification of descriptive taxanomic fields (that can fit into high-order categories). The survey already presented files, databases, registry, and hardware as categories. The fields respondents mentioned were: files (6), network packets (6), memory/memory dumps (4), application data (3), registry entries (2), type & location (2), operating system (2), data

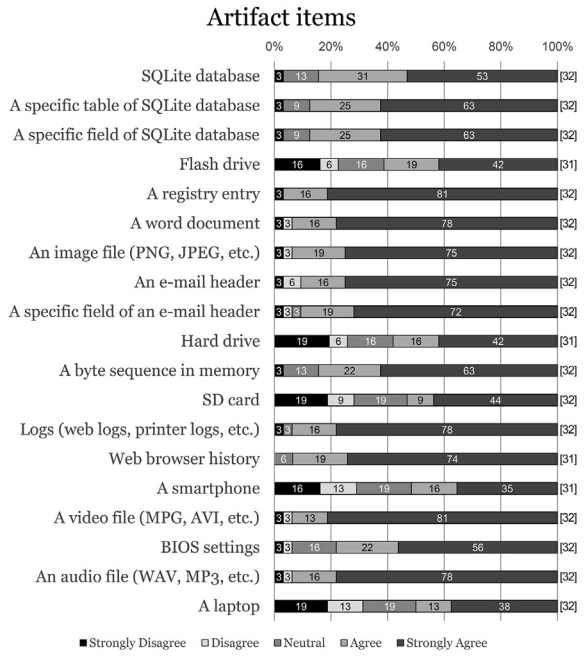


Fig. 2. Results of questions which asked respondents if the stated item should be considered a forensic artifact.

structures (2), email & webpages (2), metadata (1), sockets (1), file system (1), hashes (1), stored/volatile (1), category matches between subfields (1), external corroborating sources (1), processes (1), software (1), users (1), and device configuration (1). Some thought artifacts should be categorized by something more dynamic such as a tree (3), purpose/action (2), and physical/logical/data containers (1).

We decided that these taxanomic fields were well incorporated into CybOX and consequently should be used to improve it. Figs. 4–7 illustrate that most fields from the survey were deemed important to document by the respondents. Some of these are present in CybOX objects already; others are not and should be incorporated in the future. Furthermore, Fig. 2 (smartphone and laptop items) and Fig. 3 (hardware category) indicate that most professionals did not strongly support hardware classification as artifacts, when compared to the other responses.

Limitations

Although the sample size was large enough for the purposes of this paper, a larger sample would have been desirable. However, our size should be acceptable due to the modest size of the cyber forensics domain. Consulting sizes of organizations, forums, and groups, such as the Digital Forensics Training group on LinkedIn or the First Forensic Forum,[6] is the only current measure of the target audience at the moment. Even so, these can still be considered small since a basic search will result in much larger populations for other domains.

[6] https://www.f3.org.uk (last accessed Feb. 2, 2016).

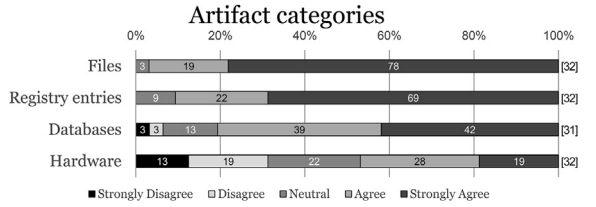


Fig. 3. Results of questions which asked respondents if the stated category should exist for digital forensics artifacts.

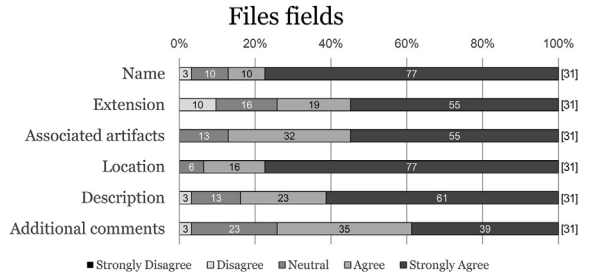


Fig. 4. Results of questions which asked respondents if the stated field should exist for describing artifacts in the Files category.

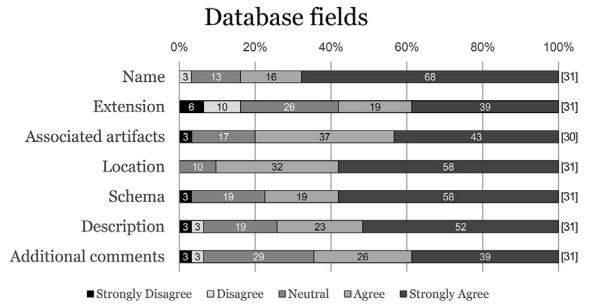


Fig. 5. Results of questions which asked respondents if the stated field should exist for describing artifacts in the Database category.

Free response questions had a wide variety of answers due to different interpretations (and misunderstandings) of the questions' abstract nature. Many answered within the context of low-level implementation and schemas. These specific answers (not the respondent) had to be disregarded (often they were reiterations of things that already existed in prior work). This could have been eliminated by stating the scope more clearly in each question, rather than simply in the disclaimer at the beginning of the survey, which most participants of surveys tend to skip. The survey design could also have been enhanced by having a "decline to respond" option to encourage response rate and minimize "missing data."

Proposed definition and model

This section describes our proposed definition, ontological model, and reiterates the importance of having a procedure like the one proposed in the Results section.

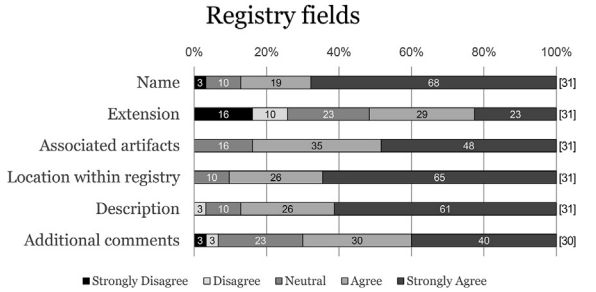


Fig. 6. Results of questions which asked respondents if the stated field should exist for describing artifacts in the Registry category.

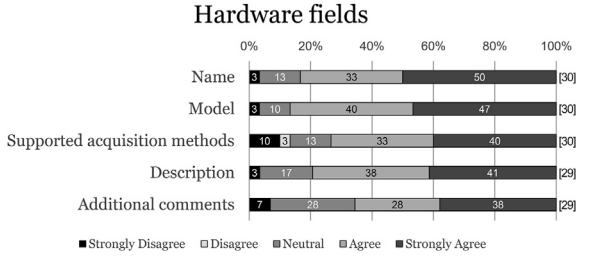


Fig. 7. Results of questions which asked respondents if the stated field should exist for describing artifacts in the Hardware category.

Definition

In creating a definition for a forensic artifact we consulted archaeology and archival science. The process of recovering, documenting, and storing objects defines items as artifacts in these domains. Consequently, we added the word "curated" to the term to make this explicitly understood by the community — Curated (digital) Forensic Artifact (CuFA).

By culminating the findings from the survey (only the top two most frequent themes from the bullet point list) presented in the Results section and the summation of previous definitions and common usages (Appendix) discussed in the Developing a definition section we propose the following stipulations for the linguistic-conceptual definition of a CuFA:

- Must be curated via a procedure which uses forensic techniques, such as the one proposed in the Results section.
- Must have a location in a useful format (when applicable).
- Must have evidentiary value in a legal proceeding.
- Must be created by an external force/artificially.
- Must have antecedent temporal relation/importance.
- Must be exceptional (based on accident, rarity, or personal interest).

Despite everything on a computer actually having a location, one must remember that the purpose of a CuFA is to find evidence on varying systems in order to improve future investigations. Therefore, location must be represented in a meaningful format that is most likely static

between different devices; the most stable/default format would be disk-related. In other words, memory location is unreliable across devices due to their allocation and runtime usage being different. Disk partitions, sectors, and other representations such as the location of user runkeys mentioned by Castle in the Developing a definition section are more likely to aid practitioners in looking for evidence across varying models and types of hardware (and maybe even different versions of the same operating system). If no useful format is possible we allow this requirement to be absent (all other requirements must still be met).

Since the definition demands the object be of "evidentiary value" we suggest this requirement be implemented with a tag/flag, which would indicate whether the specific CuFA had been successfully submitted and used in a court of law before or not. Although we found the researcher perspective to be the most common of usages, we prefer to leave this out of the formal definition since it is a consequence of usage without a standardized definition. Regardless, many of the items from the papers in the Appendix would remain identified as artifacts under the CuFA definition.

Ontological model

Based on the aforementioned definition and previous work, we established an ontological model shown in Fig. 8. The requirements from the proposed definition attempted to unify the variety of items that would be present in a CuFA database; for something of interest to be considered a CuFA it cannot be missing any of these fields (except location). Thus, items of unknown significance should be referred to as potential CuFAs or simply items of interest.

As the results of the survey made evident, location is a consistently desired piece of information, so we determined that using a high-level categorization for it would be useful; hence the *Location type* field. In addition, it will be necessary to store the location of other related CuFAs that were found for a case when making a database entry. In other words, entries should be made after investigations are finished and types of items have been established as evidence, resulting in a linked list (of pointers) of unique CuFAs that traces the leads investigators took. Although searching by location and type as Brady et al. (2014) suggested to find new potential CuFAs will still be performed, we feel this extra amount of detail will help investigators better understand the course of action at a brief glance. Possibly, this will resolve the current uncertainty of not knowing how to categorize CuFAs that act as containers for other items, because it will allow layers of abstraction to be clearly understood. For this reason we think a mandatory tag should exist for all entries to show whether the CuFA was a container (found within another CuFA).

This cannot stand on its own though. We decided that CybOX was an excellent, concrete low-level schema to help discoverers curate their artifacts when uploading them to a database such as the AGP (the AGP is currently under development; see Future work for more information). Details of CybOX's design will not be discussed but one area of improvement was identified: CybOX should involve subfields. This is often a matter of design left to the programmer that creates the system. Still, breaking up location fields into disk sector/partition, filepath, key/value pairs, and so on would help to keep interpretation consistent and comparable across platforms and agencies. Similar subfields should be present for other ambiguous fields (e.g. *Device* field in Fig. 8), and at least one of the subfields required to be completed.

Once more, location must be thought of in terms of the definition. If a CybOX object has a location-related field it may be used to satisfy the CuFA location requirement, as long as it represents a lowest-common denominator format which allows discovery of the CuFA across systems in the future. But it is not mandatory. Stating the physical location of an item only existing in memory would not aid investigation because this would differ between devices. This does not mean all CybOX objects which satisfy the location requirement have an explicit location field. The *Windows Registry Key* object type has *key* and *subkey* fields which would help locate said objects on disk for different devices. Even though this data may be copied into memory, where an investigator may retrieve it from, the lowest-common denominator format (keys and subkeys on disk) would be logged as the CuFA requirement.

The goal of this type of schema is comprehensiveness along with flexibility. Brady et al. (2014) mentioned that Encase Case Analyzer can document findings in a SQLite database, but the terminology is inconsistent. Our model could still be used alongside other models (CybOX could be replaced, accompanied, or altered) but would help standardize the items that are entered into databases, the way they are logged, and how investigators interpret the information (leveraging a more investigative viewpoint).

Conclusions

In this paper we identified requirements all items should have in order to be considered an artifact, and additionally stated that foregoing a curation process should be a new standard within the field. Thus, we suggested all items that meet these artifact requirements be named CuFAs.

Location type was introduced, centered around the creation location of an item, to incorporate this definition into an ontological model. The model cannot stand on its own though and needs to be coupled with a low-level implementation.

When comparing the results of the survey with the list of items/fields on ForensicArtifacts.com and other sources, it became clear that current models, including CybOX, are still not comprehensive enough (metadata, hashes, file systems, and operating systems were some missing fields) (Castle and Metz, 2015). We decided CybOX embodied the most comprehensive taxonomic objects/fields out of current options and offered a couple improvements (again, it could be substituted or used alongside another schema). First, dynamic fields need to be implemented in a manner which creates clear subfields, requiring the logger to choose one and input a correct format, enabling direct comparison of CuFAs. Second, using a linked-list might improve the investigative manner in which artifact

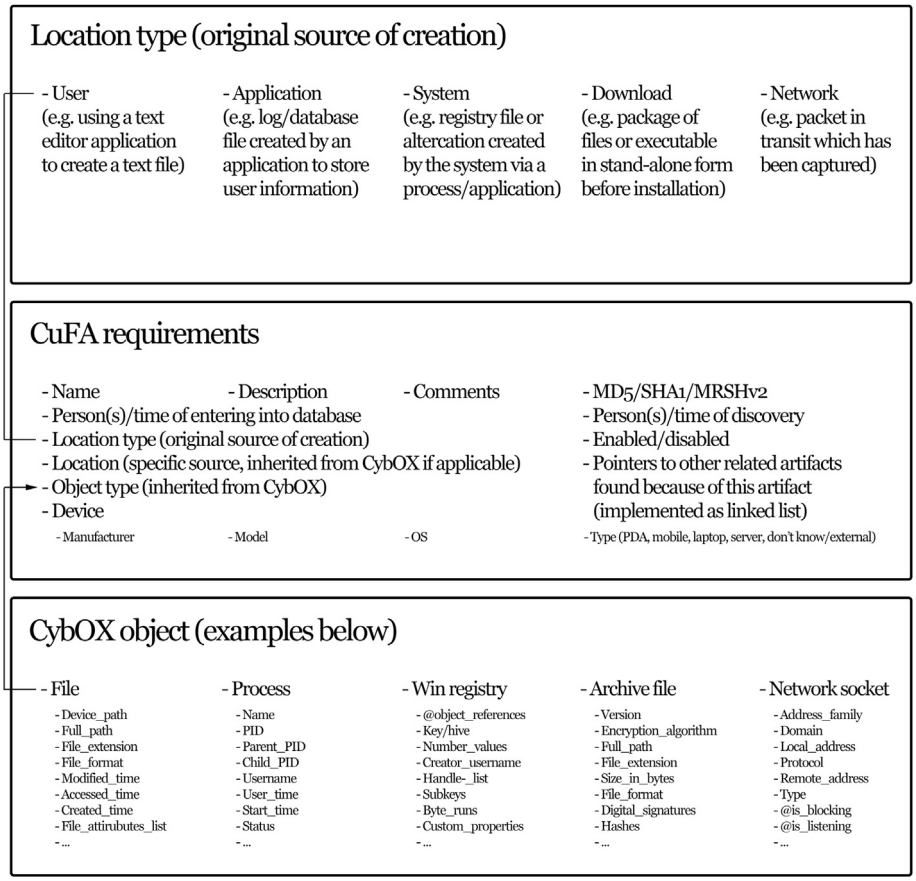


Fig. 8. The proposed ontological model uses *CybOX object* to fill specific low-level fields while the *Location type* attempts to create high-level categorization. All requirements must be met for an object to be considered a CuFA (except location; see the Ontological model section). The *Object type* requirement field at the end of the arrow illustrates inheritance from the CybOX object (beginning of arrow).

databases are used even further, by allowing the logic of an investigation to be retraced from CuFA to CuFA.

Finally, a procedure for curation was identified via survey responses, creating guidelines users of a CuFA database should use to investigate, log, and search. These contributions will increase efficiency and allow better sharing of data, and may facilitate research on "digital evolution" over time/versions.

This initial work may not be enough to create standards immediately. A more comprehensive and larger survey would solidify findings. Thus, we call for collaboration between organizations to attempt to use our findings to develop a more inclusive mechanism for creating a standardized definition. Regardless, we hope the detailed-yet-flexible nature of our model and the obvious trends in definitions/usage will drive discussion and future work to mandate standards based on these ideas.

Future work

The example of Gene Ontology given by Brady et al. (2014) supports the need for fields to be linked. Perhaps the easiest way to do this in the future would be to have two linkage boxes (in addition to a value box) for each field where one could select other fields from a dropdown menu to indicate relation. Currently, the Artifact Genome Project (AGP) is attempting to overcome some of the obstacles involved in incorporating a more standardized implementation. It is based on our proposed definition and model, and utilizes elements of CybOX (many of the CybOX objects have already been discarded due to absorption by other re-defined objects, lack of utility in forensics, and over-specificity). The AGP will attempt to create a repository of CuFAs available publicly for researchers, especially, and practitioners to log CuFAs and track investigations via the CuFA linked-lists (effective techniques for recovering, copying, quarantining, and so on may also be logged, in accordance with the procedure presented in the Results). Although many tools, such as GRR, exist to acquire and analyze artifacts the primary point of the AGP will be to create a centralized database regardless of type of tools used in the process. It will also help to standardize the way people upload and categorize CuFAs, helping to reduce wasted time searching for desired types of items. One way it will do this is to pull up possible CybOX objects based on the fields users enter, and then allow users to flag the object they select for different operating systems and levels of legality (has the object officially been used in a court of law); selecting inappropriate object types will become less common.

Accompanying this advancement should also be tools to facilitate the use of such databases. OSXAuditor was one example: it helped to automatically search locations on a running system/system image to find items of interest, extract them, and verify the reputation of files using VirusTotal, MHR, and Malware.lu (or your own database). It could also aggregate logs from locations and put them into a zipball. The final results could be rendered as text or HTML and sent to a Syslog server (Roberts, 2013). OSXAuditor is no longer maintained and has been set aside for OSXCollector, an automated toolkit oriented towards enabling analysts to answer questions, like how malware got onto an infected computer, quicker. Its single Python file creates a package of the collection (outputed to JSON) and useful files like system logs to pass off to analysts who can look through information on startup, quarantines, operating system info, browsers, downloads, kernel extensions, file timestamps, etcetera (Yelp, 2016). Tools such as this will be necessary in the future to expedite using CuFA databases and allow investigators to focus on higher levels of abstraction (at least when desired).

Acknowledgments

We would like to thank Jason Moore for the survey design & testing, as well as Kyle Anthony and Devon Clark for helping in the design of the ontological model based on the data & code they were working on for the AGP. The AGP and the material presented in this paper is supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI001-05.

Appendix

For a full explanation of this Appendix see the Perspectives and usage section.

Items	Category	Paper & perspective
User credentials, personal details, activities, location; Activity timestamps;	Databases	Azfar et al., 2015 Researcher
Images;	Media	
Opened/saved files; Email attachments; Skype log (chat & transfer); Index.dat (downloads);	File download	Goh, 2014 Researcher
User assist (program launch); Last executed files by app; Run command executed; App compatibility cache; Taskbar jump list; Prefetch/service event logs;	Program execution	
Opened/saved fields; Last executed files by app; Recently opened files; Shellbags; Shortcut files (LNK); Taskbar jump list; Prefetch files; IE history files;	Files created & opened timeline	
Search assistant/history; Keywords search from Start Menu; Last executed files by app; Hidden files in dir (Thumbs.db); Recycle bin; IE history files;	Deleted files	
Current system timezone; Network history; IE cookies; Time website visited;	Physical location	
USB key identification; USB device plug & play times; GUID of mounted devices; Volume serial number; Drive letter & volume name; Shortcut link files (LNK); Plug & play event log;	Drive usage	
Last password change; Last login; Successful/failed login; Login types for account; Remote desktop usage;	Account usage	
IE history; IE cookies; IE cache files for web content; Automatic crash recovery; Local stored object & flash; Network history;	Browser usage	
Contact details & profile; Picture URLs; Photo uploads; Comments posted; Timestamps; Previously logged in users; Friends with active chat; Created albums; Pictures viewed with app; Mailbox/chat messages;	Facebook	Mutawa et al., 2012 Researcher
User names; Profile picture URLs; Tweets posted; Other activity (e.g. device);	Twitter	
Usernames/passwords; Post comments; Timestamps; Cookies & cache files;	MySpace	
Local folder; Metro apps; IE10 websites visited; Journal notes; Desktop tools; Metro app web cache; Metro app cookies; Cache; Cookies; Microsoft folder; Digital certificates; User contacts; Application settings;	Windows	Thomson, 2012 Researcher
Ntuser.dat; SAM; System; USB storage devices; Software;	Registry	
Space carved from SSD; EFI-system objects from carving; Grub in boot sector;	Chrome operating system	Corbin, 2014 Researcher
User directory of Chrome files; Google website history; Bookmarks; Cookies; Download, search, login history; Most visited websites; Cache;	Contained/inner items	
Temporary content; User content; System support; System updates; File timestamps;	Xbox One NTFS partitions	Moore et al., 2014* Researcher
Bit assignments; Browsing content records; Database files; Page files; Log files;	Private browsing	Chivers, 2014 Researcher

(*continued*)

Items	Category	Paper & perspective
Registry; Application data folder;	Google client-side	Gupta et al., 2013 Researcher
Keyword searches; Usernames/passwords; Most recently used/cache data;	Registry	Mee et al., 2006 Researcher
USB device database;	Database	Collie, 2013 Researcher
Text; Images; Sketches; Videos; Location data; Audio; Video;	Smartphone network traffic	Walnycky et al., 2015 Researcher
Chat logs; User info in SQLite files;	Local	
Application settings; Installation paths; Program compatibility assistant; Magnetic/registry key links; IE integration; Statistics; Open with list; Windows routing service tracing; Remote access service tracing; File associations; Uninstallations;	Registry directory & key/value Other BitTorrent association	Lallie and Briggs, 2011* Researcher
Apple system log; Crash reporter; Diagnostic messages; FSEvents API; Preference settings; Saved application state; Spotlight; Swap files/paging/cache; Temporary data;	OSX	Sandvik, 2013* Researcher
Prefetch; Thumbnail cache; Paging file; Registry; Windows search; Bash history; GVFS virtual file system; Recently used; X session manager;	Windows Linux	
Install path (install/delete); Registry keys (install/delete); Prefetch files (install/delete/runtime); VDF signatures;	Virtual disk encryption tool	Lim et al., 2010 Researcher
User/attacker geoIP, source/private IP, SIP user agent, device, habits; Not found 401, 404; Options method; Frame time; Source IP address; Destination IP address; SIP from/to; SIP contact; SIP user agent via call-ID; Cseq; SDP owner, connection, session name, media attributes; Info request/response;	Network traffic SIP/SDP header	Psaroudakis et al., 2014* Researcher
RAM; Swap files; Registry;	Accelerator Plus	Yasin et al., 2009 Researcher
Proxy settings; History of downloaded files; Files requested to download; Incomplete downloaded files; Password protected websites; Site grabber; Uninstall location; Downloaded files; Site grabber; Uninstall process; Encrypted password storage;	Registry Log files	Yasin et al., 2010 Researcher
Log analysis; RAM analysis;	Digsby messaging client	Yasin and Abulaish, 2013 Researcher
Registry keys/values; Directories & files;	Steganography	Zax and Adelstein, 2009 Researcher
Antivirus/quarantine-related; Authentication; Web browser; Configuration/registry files; Containers for execution events; External media data/events; Log files; Memory/volatile data; Networking state; Running processes; Installed software; System-related; User-related file/type/location;	GRR & ForensicArtifacts.com	Castle and Metz, 2015 Collector
Autorun locations; System preferences; System settings & info; Sleep/hibernate/swap image file; Kernel extension; Software installation; Miscellaneous system info; Networking; Autorun locations; Users; User directories; Preferences; Logs; User accounts; iDevice backup; Recent items; Miscellaneous; iCloud; Skype; Safari; Firefox; Chrome; Mail;	System User Application	Stirparo, 2015* Collector
File downloads; Program execution; File opening/creation; Deleted file/file knowledge; File physical location; USB/drive usage; Account usage; Browser usage;	SANS cheat sheet	Lee, 2015 Collector

(*continued on next page*)

(*continued*)

Items	Category	Paper & perspective
File; Process; Win registry key; Win service; Win thread; Archive file; Mutex; URI; Domain name, address, & hostname; Port; Network socket; Link; DNS record; ARP cache; URL history; Email message; Socket address; Pipe; Win mailslot; Win memory page region; Win filemapping; Semaphore; Win event; Win critical section; Win handle; WHOIS;	CybOX objects	MITRE Corporation, 2015 Collector
Apple serial number ID; Mobile phone handset ID; Network address ID; SIM card ID; USB device ID;	Device identifier	Brady et al., 2014 Collector
File system ID; IP address ID; SSID;	Logical identifier	
IP addresses/domains; Mutexes; Open(ed) files; Services; Registry keys/values/write times; System date; Process names/timestamps; Thread/network timestamps; UserAssist last run times;	Volatility Framework	Levy, 2011 Tool
Event logs in XP; PE timestamps;	Volatility Timeliner	
PDF, TXT, RTF, Office, etc. files;	Document files	MAGNET, 2015 Tool
USB devices; File system info; Network share info; Link files (shortcuts); User accounts; Startup items; OS info; Shellbags; JUmplists; Event logs; Prefetch files; Timezone info;	Windows	
Outlook web app & email client; Microsoft sharepoint; Mbox email;	Corporate email	
Microsoft Lync/OCS;	Instant messaging	
Calendar; Call logs; Contacts; iMessage/SMS; Native notes;	iOS backup	
SMS & voicemail; Browser; Cell.cache & Wifi.cache; Maps; Pictures; Notes; Contacts & call logs; Downloads; Email; Application snapshots; iOS owner info, notes, wifi/Bluetooth info, user word dictionary, spotlight searches, calendar events, installed applications;	Phone apps	
Network connections; Running processes; Connected network shares/drives; Alert on remote connections; Network interfaces; Logged on users; Scheduled tasks; Services;	Triage	
Instant messaging chats; Media; P2P file sharing; Social networking sites; Webmail applications; Web-related activities; Webpage recovery; Xbox;	Internet	

References

Azfar A, Choo K-KR, Liu L. Forensic taxonomy of popular android mHealth apps. 2015. Technical Report University of South Australia.

Bariki H, Hashmi M, Baggili I. Defining a standard for reporting digital evidence items in computer forensic tools. In: Digital forensics and cyber crime. Springer; 2011. p. 78–95.

Brady O, Overill R, Keppens J. Addressing the increasing volume and variety of digital evidence using an ontology. In: Intelligence and security informatics conference (JISIC), 2014 IEEE joint; 2014. p. 176–83.

Brinson A, Robinson A, Rogers M. A cyber forensics ontology: creating a new approach to studying cyber forensics. Digit Investig 2006;3:37–43.

Casey E, Back G, Barnum S. Leveraging cybox to standardize representation and exchange of digital forensic information. Digit Investig 2015;12:S102–10.

Castle G. Black hat usa 2014-forensics: Grr find all the badness, collect all the things. 2014.

Castle G. Grr artifacts. Blackhat; 2014b.

Castle G, Metz J. Forensic artifact labels. 2015.

Chivers H. Private browsing: a window of forensic opportunity. Digit Investig 2014;11:20–9.

Collie J. The windows iconcache.db: a resource for forensic artifacts from {USB} connectable devices. Digit Investig 2013;9:200–10.

Corbin G. The Google Chrome operating system forensic artifacts [Ph.D. thesis]. Utica College; 2014.

Dictionary.com. Definition of artifact. 2015.

Dietrich D, Adelstein F. Archival science, digital forensics, and new media art. Digit Investig 2015;14:S137–45. The Proceedings of the 15th Annual {DFRWS} Conference.

Garfinkel S. Digital forensics xml and the dfxml toolset. Digit Investig 2012;8:161–74.

Goh T. Challenges in Windows 8 operating system for digital forensic investigations (Doctoral dissertation, Auckland University of Technology). 2014.

Gupta A, Verma R, Gupta G. Client side forensics investigation of google services. In: IEEE symposium on security and privacy; 2013.

Lallie HS, Briggs PJ. Windows 7 registry forensic evidence created by three popular bittorrent clients. Digit Investig 2011;7:127–34.

Lee R. Sans digital forensics and incident response poster. 2015.

Levy J. Time is on my side. 2011.

Lim S, Park J, Lim K-s, Lee C, Lee S. Forensic artifacts left by virtual disk encryption tools. In: Human-centric computing (HumanCom), 2010 3rd international conference on; 2010. p. 1–6.

MAGNET. Artifact lists. 2015.

Malafsky GP, Newman BD. Organizing knowledge with ontologies and taxonomies. 2009. Technical Report TECHi2.

Mee V, Tryfonas T, Sutherland I. The windows registry as a forensic artefact: illustrating evidence collection for internet usage. Digit Investig 2006;3:166–73.

Merriam-Webster Dictionary. Definition of artifact. 2015.

MITRE Corporation. Cyber observable expression (cybox). Github; 2015.

Moore J, Baggili I, Marrington A, Rodrigues A. Preliminary forensic analysis of the xbox one. Digit Investig 2014;11(Suppl. 2):S57–65. Fourteenth Annual {DFRWS} Conference.

Mutawa NA, Baggili I, Marrington A. Forensic analysis of social networking applications on mobile devices. Digit Investig 2012;9(Suppl.):S24–33. The Proceedings of the Twelfth Annual {DFRWS} Conference12th Annual Digital Forensics Research Conference.

Oxford Dictionaries. Definition of artifact. 2015.

Poli R, Healy M, Kameas A. Theory and applications of ontology: computer applications. Springer; 2010.

Psaroudakis I, Katos V, Saragiotis P, Mitrou L. A method for forensic artefact collection, analysis and incident response in environments running session initiation protocol and session description protocol. Int J Electron Secur Digit Forensic 2014;6:241–67.

Roberts SJ. Osx auditor. Github; 2013.

Sandvik R. Forensic analysis of the tor browser bundle on OS X, Linux, and Windows. 2013. Technical Report Tor Tech Report.

Stirparo P. Mac4n6 osx and ios artifact collection. 2015.

SWGDE/SWGIT. Scientific working groups on digital evidence and imaging Technology. 2nd ed. 2015.

Thomson A. Windows 8 forensic guide. 2012. Technical Report The George Washington University.

Walnycky D, Baggili I, Marrington A, Moore J, Breitinger F. Network and device forensic analysis of android social-messaging applications.

Digit Investig 2015;14(Suppl. 1):S77–84. The Proceedings of the Fifteenth Annual {DFRWS} Conference.
Yasin M, Abulaish M. Digla – a digsby log analysis tool to identify forensic artifacts. Digit Investig 2013;9:222–34.
Yasin M, Cheema AR, Kausar F. Analysis of internet download manager for collection of digital forensic artefacts. Digit Investig 2010;7:90–4.
Yasin M, Wahla M, Kausar F. Analysis of download accelerator plus (dap) for forensic artefacts. In: IT security incident management and IT forensics, 2009. IMF '09. Fifth international conference on; 2009. p. 142–52.
Yelp. Osx collector. Github; 2016.
Zax R, Adelstein F. Faust: forensic artifacts of uninstalled steganography tools. Digit Investig 2009;6:25–38.