Digital Investigation 26 (2018) S47–S58

Contents lists available at ScienceDirect

Digital Investigation

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

DFRWS 2018 USA — Proceedings of the Eighteenth Annual DFRWS USA

Experience constructing the Artifact Genome Project (AGP): Managing the domain's knowledge one artifact at a time

Cinthya Grajeda [1], Laura Sanchez [2], Ibrahim Baggili*, [3], Devon Clark [4], Frank Breitinger [5]

Cyber Forensics Research and Education Group (UNHcFREG), Tagliatela College of Engineering, ECECS, University of New Haven, 300 Boston Post Rd., West Haven, CT, 06516, United States

Keywords:
Forensics
Artifacts
Applications
Education

A B S T R A C T

While various tools have been created to assist the digital forensics community with acquiring, processing, and organizing evidence and indicating the existence of artifacts, very few attempts have been made to establish a centralized system for archiving artifacts. The Artifact Genome Project (AGP) has aimed to create the largest vetted and freely available digital forensics repository for Curated Forensic Artifacts (CuFAs). This paper details the experience of building, implementing, and maintaining such a system by sharing design decisions, lessons learned, and future work. We also discuss the impact of AGP in both the professional and academic realms of digital forensics. Our work shows promise in the digital forensics academic community to champion the effort in curating digital forensic artifacts by integrating AGP into courses, research endeavors, and collaborative projects.

1. Introduction

Digital Forensics is a multidisciplinary domain that involves computing, law, criminology, psychology and other disciplines. At the core of the domain, however, is the Acquisition, Authentication and Analysis (AAA) of digital evidence. In the real world, practitioners typically find data of forensic value in digital forensic artifacts. The Scientific Working Group on Digital Evidence (SWGDE) defines an artifact as *"Information or data created as a result of the use of an electronic device that shows past activity"* (SWGDE, 2015).

As technology continues to evolve and consumption of technological goods and services continue to grow, it becomes increasingly important to maintain current knowledge regarding digital artifacts created by systems (Garfinkel, 2013). This includes understanding what artifacts look like and where to find them. While automated tools such as Encase[6] and FTK[7] exist to aid the digital forensics community in obtaining, deciphering, organizing, and storing evidence collected during an investigation, these tools do not explicitly provide in-depth information about the makeup of artifacts; they typically only serve to indicate that artifacts potentially exist and they decode them if a decoder is present. Current tools do not contribute to establishing and maintaining a systematic approach for artifact knowledge management.

This work aims to help alleviate the following challenges faced by the digital forensics community:

- At the time of writing, there was no vetted, current, curated, publicly available, crowd-sourced, systematic approach for digital forensics artifact curation.
- Many times, practitioners around the world work in isolated environments and have to repeat the reverse engineering and decoding of artifacts of relevance to their work wasting time, money and resources.
- The community does not have a mechanism for studying the basic scientific principles of what a digital forensics artifact is, and how it might change over time.
- It is a difficult problem to keep up with artifacts found on devices with different operating systems and applications.

* Corresponding author.
E-mail addresses: Cgraj1@unh.newhaven.edu (C. Grajeda), lsanc3@unh.newhaven.edu (L. Sanchez), IBaggili@newhaven.edu (I. Baggili), Dclar4@unh.newhaven.edu (D. Clark), FBreitinger@newhaven.edu (F. Breitinger).

[1] http://www.unhcfreg.com/.
[2] http://www.unhcfreg.com/.
[3] http://www.Baggili.com/.
[4] http://www.unhcfreg.com.
[5] http://www.FBreitinger.de.
[6] https://www.guidancesoftware.com/encase-forensic?cmpid=nav_r.
[7] https://accessdata.com/products-services/forensic-toolkit-ftk.

https://doi.org/10.1016/j.diin.2018.04.021

To tackle the aforementioned challenges, we conceived the Artifact Genome Project (AGP).[8] Initiated in 2014 and launched in 2017, AGP aimed at becoming a unified digital forensic artifact curation platform. AGP can be employed by practitioners and researchers for educational and investigative purposes. Similar to the Human Genome Project (HGP) "whose goal was the complete mapping and understanding of all the genes of human beings" (NHGRI, 2016), AGP aspires to create a fundamental map of digital forensic artifacts by indicating their type and mapping out their unique and identifiable characteristics, as well as their location. Our primary contributions from this work are as follows:

- Our work has resulted in the largest, vetted, freely available digital forensics artifact platform.
- Our work serves as the primary implementation of the Curated Forensics Artifact (CuFA) model by Harichandran et al. (2016).
- Our work has catalyzed a community, crowd-sourcing based model for artifact collection.
- We share our design choices, lessons learned and experience from building and maintaining such as system.

This paper is organized as follows. Section 2, presents background information and related work, including previous research. Next, Section 3 and functionality, details the design and functionalities of AGP. The impact of AGP in the professional realm and in academia is discussed in Section 4, followed by a discussion in Section 5 of a few of the artifacts in AGP. Section 6 discusses the collection of data from research queries and how the data is used, while Section 7 discusses our experiences in building AGP and what we learned. Finally, Section 8 concludes the paper and Section 9 presents future work.

2. Related work

2.1. Forensic artifact analysis

There is an incredibly large body of literature in digital forensics science that identifies and analyzes artifacts of forensic value on a variety of systems, such as mobile devices (Bader and Baggili, 2010; Al Marzougy et al., 2012; Iqbal et al., 2013), Supervisory Control and Data Acquisition (SCADA) (Denton et al., 2017; Senthivel et al., 2017; Ahmed et al., 2017), smart watches (Baggili et al., 2015; Ricci et al., 2016), cloud storage forensics (Hale, 2013; Quick and Choo, 2014; Roussev and McCulley, 2016; Roussev et al., 2016), drones (Clark et al., 2017), and mobile and desktop applications (Al Mutawa et al., 2012; Walnycky et al., 2015; Zhang et al., 2017a; Al Mutawa et al., 2011; Marrington et al., 2012). These works represent a small fraction of published literature in this domain. To go over the entire body of knowledge in artifact analysis is beyond this paper's scope.

2.2. Schemas and ontologies

Before being able to talk about digital forensic artifacts, we need a "standardized language for encoding and communicating high-fidelity information about cyber observables" (MITRE, 2014). CybOX is one solution, which is a language with a list of required and optional attributes for each object type, i.e., a file object, and it is mainly classified by source. Moreover, it assigns a Globally Unique Identifier (GUID) to the object, making it unique and easily searchable in a database. CyBOX allows for the recording of the state of a system and logging differences based on timestamps or similarity digest and has been adopted by the cybersecurity community due to its open source nature and precise classification scheme for objects. It is also used, in part, in our AGP implementation. There are also other high-level ontologies used in the community, such as Structured Threat Information eXpression (STIX).

STIX also uses the low-level CybOX schema. This exhibits details of objects of interest and allows it to be identified and categorized in a forensically sound manner. For example, malicious IPs could be identified as threat actors. Its successor, the Digital Forensics Analysis eXpression (DFAX), is a more advanced system used with CybOX to collect and organize even more process details, such as that of a chain of custody. Finally, the Unified Cyber Ontology (UCO) attempts to collect and unify common cyber domain objects that are even compatible with DFAX and STIX (Casey et al., 2015).

2.3. Attempts at an artifact database

Two main attempts were implemented for archiving and sharing artifacts with the digital forensics community.

ForensicArtifacts.com was an initial attempt at trying to build a forensics community-sourced artifact repository. The oldest uploaded artifact goes back to 2010. According to the website, it "was built to become a repository for useful information forensic examiners may need to reference during the course of their analysis" (Forensic Artifacts, 2012). While this might have been true, the platform appears to be outdated, with little structure to the uploaded artifacts. Unfortunately, this is a challenge the digital forensics community faces with open source artifact or dataset repositories. If there are limited resources allocated to maintain and keep the repository up-to-date, they start to become obsolete. A novel incentive this platform offered, however, is a SANS[9] Lethal Forensicator Coin for users that submit six or more artifacts or Indicators of Compromise (IOC) in a year.

Unlike AGP, ForensicArtifacts.com does not require users to subscribe or pass a vetting process to submit artifacts. Additionally, the site is rather modest and offers only one submission form for any type of artifact. It is also not clear if all artifacts go through a sanitation process to ensure artifact integrity and efficacy.

In early 2017, Magnet Forensics introduced the *Artifact Exchange* (MagnetForensics, 2017). Available to the digital forensics community, the Artifact Exchange allows practitioners to upload or download custom artifacts. Artifacts are built in XML or Python and can be integrated into Magnet Forensics' AXIOM – a digital investigation platform for smartphones, computers, and the cloud. While available to the community-at-large, because the artifacts are built using AXIOM's API, the artifacts will work only with their product. At the time of it's launch, the Artifact Exchange only had 50 applications available.

2.4. Artifact related systems

Bhoedjang et al. (2012) described the process of creating and implementing an automated digital forensic tool, XML-based indexing and querying for digital forensics (XIRAF), to collect and organize artifacts so they may be efficiently searched. Utilized by law enforcement in the Netherlands, XIRAF was introduced as a service to allow investigative teams to access artifacts and work collaboratively regardless of technical skills. The automated processing of artifacts "removes the requirement on the investigator to know about the technicalities of the different systems and

[8] https://agp.newhaven.edu.

[9] This is a private U.S. for-profit institute that specializes in information security and cybersecurity training. For details, see https://digital-forensics.sans.org/. Last accessed: 1/20/2018.

applications that can produce a certain type of digital artifact" (Bhoedjang et al., 2012).

Succeeding XIRAF, Hansken was developed as a solution for processing big data in a more efficient manner. The design of Hansken was influenced by three forensic drivers and several design principles (Van Beek et al., 2015). The three drivers include minimizing case lead time, making available traces within the first 48 h of an investigation; maximizing coverage, or the ability of tools to process various types of artifacts; and specializing people, a concept in which individuals have specific job duties to fulfill based on their position. Design principles were defined by considering the particular risks of implementing a big data platform. The principles included security, privacy, transparency, multi tenancy, future proof, data retention, reliability, and high availability.

3. AGP system design and functionality

At the time of writing, there was no practical, thoroughly researched, and freely available approach that combined a comprehensive standardized artifact definition, ontology and technical schema combined with CybOX to provide the benefits and capabilities that AGP does to the forensic community in a usable system.

AGP is a crowd-sourcing initiative in which digital forensics professionals conduct research and share results relating to digital artifacts. In an online system promoting a community, users can upload their findings and view the findings of others. The following sections provide details regarding the system architecture, user and administrator functionalities as well as the process for accessing and contributing to the growing system.

3.1. Curated (digital) forensic artifact (CuFA)

An essential component of AGP is work from Harichandran et al. (2016), which aimed at understanding the challenges of not having a standardized linguistic definition and ontological model for artifacts. The research consisted of distributing a survey to researchers and practitioners to learn their perceptions regarding artifacts. The data gathered, and extended research, helped explore various artifact-related models to gain a deeper understanding of what an artifact is, as well as the process it would take to sanitize artifacts. Overall, the preliminary work made the following contributions:

1. Proposed a more concrete, unified linguistic definition, and assigned it a new name: Curated (digital) Forensic Artifact (CuFA).
2. Using the survey responses and the proposed definition, an ontological model was designed for the curation of artifacts. This 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 Cyber Observable eXpression (CybOX) resulting in a searchable database organized by dynamic, taxonomic fields and tags/flags.

The model from that work is adopted by AGP (Fig. 1).

3.2. AGP architecture

AGP's system architecture is designed to take advantage of several open source software and web development platforms. This helps the system run smoothly with reduced maintenance and allows the handling of a scalable workload without reinventing the wheel. Fig. 2 shows a high-level diagram of the system architecture.

AGP runs on an Ubuntu server. PostgreSQL[10] is installed as well and is the main database system used to store data, such as account information, activity logs, and artifact data. AGP was also developed with Django.[11] The system architecture is depicted in Fig. 2 and is modular in nature. Main applications include the admin, which handles various actions such as flagging artifacts for processing, and the base, which handles interactions with the base AGP site and includes graphs and featured artifacts on the Home Page. Finally, the artifacts application handles the uploading, editing, and querying of artifacts. Search indexing and querying is managed with the Solr[12] search engine.

3.3. Functionality and features

While the main focus of AGP is to interact with the artifact database, users can also contact administrators, view artifact statistics, and interact with other users. To interact with the artifact database, users are required to log in, which brings them to a landing page (screenshot provided in the Appendix Figure B.7). From there, a user has the ability to access and perform three different actions: creating artifacts (Section 3.3.1), searching for artifacts (Section 3.3.2) and accessing *my artifacts* (Section 3.3.3).

3.3.1. Creating artifacts

As listed in Table 1, AGP has nineteen artifact default template types that a user may select from when creating an artifact. Additionally, AGP has the capability of integrating new artifact type templates if needed in the future.

The structure of each artifact template form includes common and unique attributes about each artifact type (screenshot provided in the Appendix Figure B.8). For instance, all artifact forms contain input fields for the artifact's title, which must be a unique identifier, as well as the artifact's complete directory path (where the artifact was found on a device) and the artifact's hash value (MD5 or SHA-1). Fields also exist for details unique to a particular type of artifact, such as Network Packets or Windows Registry artifacts. The form for network packets, for example, contains a field to input the raw packet hex data, while the Window's registry form contains fields for inputting key and subkey values.

Additionally, users can add three types of tags to each artifact. Users can add search tags, which is a list of keywords related to the artifact (i.e., Android). Artifact tags can also be added, which link to similar or related AGP approved artifacts in the database. Finally, users can tag other AGP users by choosing them from a list of taggable usernames.

3.3.1.1. Granular artifact tagging. In addition, AGP goes a step further and provides an advanced file tagging mechanism for a more granular approach towards understanding artifact structures. This approach allows users to highlight and label forensically relevant elements in the unique files that they have uploaded as part of the artifact. For instance, Fig. 3 displays a snippet of a SQLite database table called "convos". This type of taggable database file contains three forms of tagging. One is by table, where the user can

[10] This is an open source object-relational database management system. For details, see https://www.postgresql.org/. Last accessed: 01/11/2018.

[11] This is an open source web application framework, written in Python. For details, see https://www.djangoproject.com/ web framework. Last accessed: 01/11/2018.

[12] This is an open source search platform, written in Java, from the Apache Lucene project. For details, see http://lucene.apache.org/solr/. Last accessed: 01/11/2018.

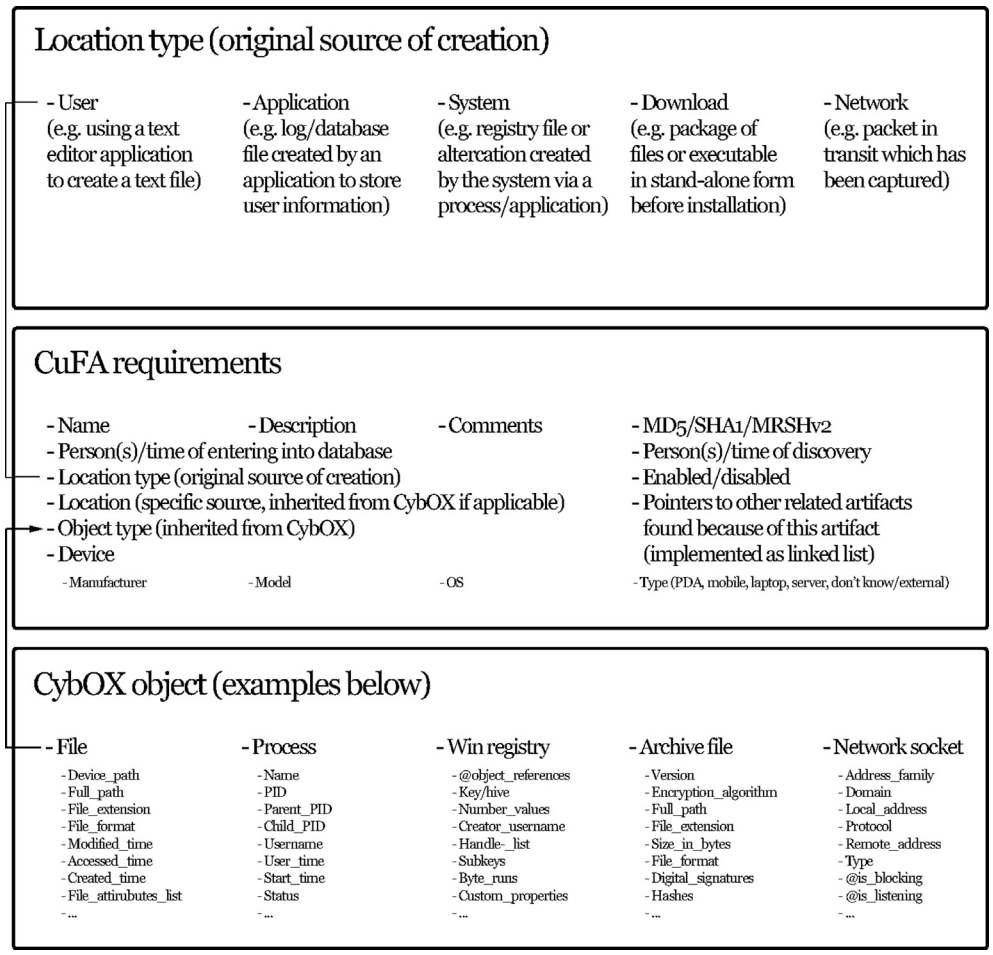


Fig. 1. The model uses *CybOX object* to fill specific low-level fields while the *Location type* attempts to create high-level categorization. All requirements must be met for an object to be considered a CuFA (except location). The *Object type* requirement field at the end of the arrow illustrates inheritance from the CybOX object (beginning of arrow) (Harichandran et al., 2016).

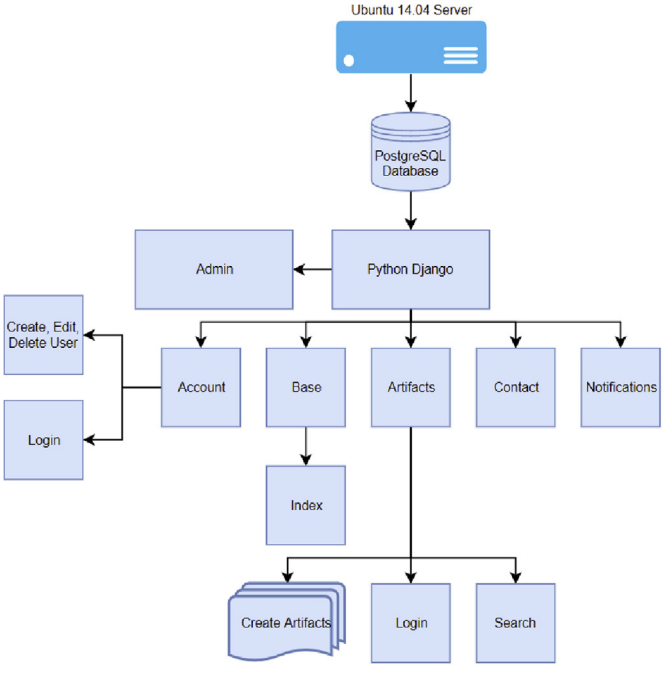


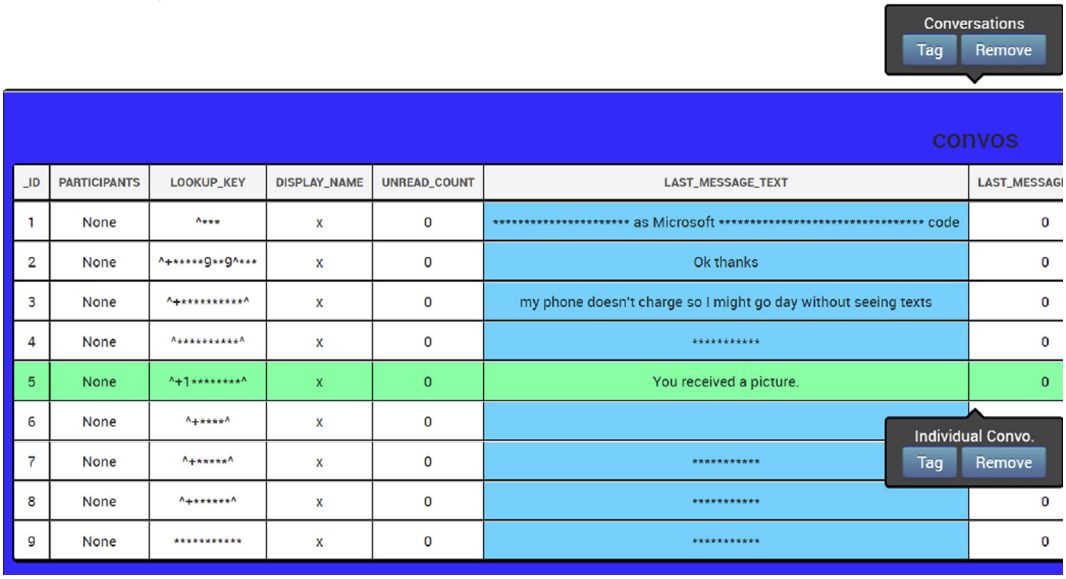
Fig. 2. AGP system architecture.

click on the outer table area and then select from a list of tags or type in their own tag. This action highlights the whole table with a blue color. Similarly, rows and columns can also be tagged within the table.

Databases are not the only type of files that can be tagged in a granular manner. AGP supports nine other different types of taggable file formats, such as text, logs, and XML. In those types of files the user can highlight certain pieces of information and add a tag (e.g., E-mail) to it. The objective of this efficient system is to assist users by rapidly shifting the user's attention to the highlighted areas of the form to locate forensically relevant structures inside the uploaded file. Users will be able to quickly spot information that might be relevant to their case.

Table 1
Default artifact object templates.

File	SMS Message	Window Registry
Disk	Network Socket	User Account
Process	Disk Partition	User Session
Memory	E-Mail Message	Volume
Code	Linux Package	Windows Event Log
Account	Network Packet	X509 Certificate
Address		



_ID	PARTICIPANTS	LOOKUP_KEY	DISPLAY_NAME	UNREAD_COUNT	LAST_MESSAGE_TEXT	LAST_MESSAGI
1	None	^***	x	0	********** as Microsoft ********** code	0
2	None	^+******9**9^***	x	0	Ok thanks	0
3	None	^+**********^	x	0	my phone doesn't charge so I might go day without seeing texts	0
4	None	^**********^	x	0	**********	0
5	None	^+1********^	x	0	You received a picture.	0
6	None	^+****^	x	0		
7	None	^+*****^	x	0	**********	
8	None	^+*******^	x	0	**********	0
9	None	**********	x	0	**********	0

Fig. 3. AGP SQLite database tagging.

3.3.2. *Searching for artifacts*

The user may search for any artifacts that have been entered into the system by all users and labeled as approved by an administrator. Users may search for related terms by entering them in the search field, or they can refine their search by using the Advanced Search feature. For an overview of the most searched keywords in AGP, see the word cloud in Fig. 6. In advanced search, users may search by specific artifact types, tag, device, date and time ranges, users, and more. Artifact results are listed by name, type, and submission date. Selecting a specific artifact will display it in read-only format. Moreover, users have the ability to download any example files attached to the selected artifact and even export the artifacts in Comma Separated Values (CSV) format organized by artifact type.

3.3.3. *My artifacts*

Finally, AGP allows users to consolidate and search through all the artifacts that they have uploaded to the database. This is similar to a traditional Search, however, it will only search for and display artifacts uploaded by the authenticated user performing the search. Users have the option of editing every artifact they have uploaded to AGP. One of the main reasons users might have to modify any of their artifacts is if the artifact has been flagged by the administrator, for instance. The user can choose to narrow their search results to artifacts with one of five particular status values, such as Queued (waiting for admin review) or Flagged (waiting for user update). Any notes from an administrator about artifacts that need to be revised, or corrected, will also be displayed to the user.

3.4. *Vetting processes*

To ensure the integrity of AGP, two vetting processes are employed.

3.4.1. *User Vetting Process*

AGP users stem from various backgrounds, including academia, law enforcement, and private and public-sector organizations. AGP is open to community members that request access to use it, however, not all applicants may meet our vetting standards. The application vetting process (Fig. 4) prevents system contamination and assures that only experts, scientists, and professionals have access to this community initiative. Therefore, every applicant goes

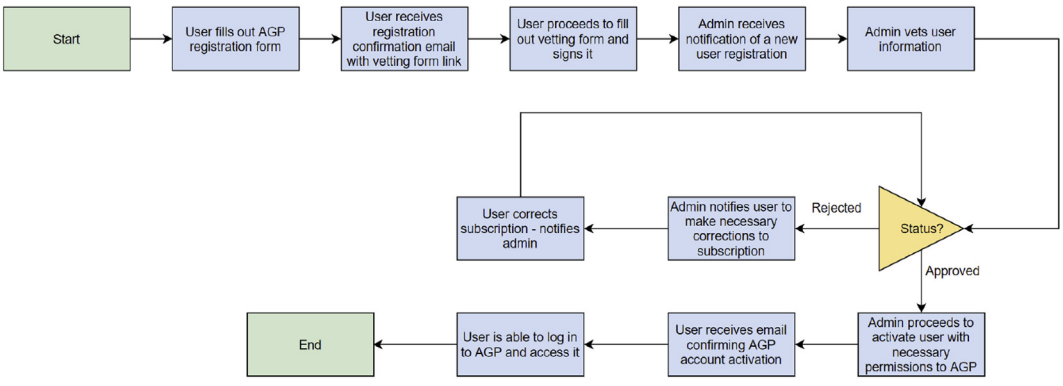


Fig. 4. AGP user vetting process.

through a vetting process in which all the information they have provided is verified.

Some of this information includes their name, username, organization, organization address and e-mail address. We explicitly ask users to apply with a professional e-mail, when possible, because it can expedite the application process. The organization the applicant belongs to is either verified via the e-mail address the user has provided or by the AGP administrator contacting applicants on an individual basis. Additionally, this information is verified through social media if it exists, especially LinkedIn accounts. Lastly, applicants are required to agree to the AGP user policy by signing the vetting form. User accounts will not perform as intended without administrator intervention and no account is made active unless it passes the vetting process.

3.4.2. Artifact Vetting Process

Active users may submit any artifact to AGP that is not already present in the system. However, artifacts are not visible to other users until they have been inspected and sanitized by AGP administrators. The vetting process of artifacts can vary, depending on the type of submitted artifact. In order for any artifact to pass the vetting process, certain conditions have to be met. Initially, when a user uploads a new artifact to AGP, the artifact status is automatically set to the queued state (See Fig. 5 for a high-level presentation of this process). Artifacts in the queue are initially assigned to different admin members for review through an automated round robin system. This approach ensures that the workload of vetting artifacts is evenly distributed and balanced between admin members. Thus, for new artifact uploads, the admin member with the least workload will get that artifact assigned to their list for review.

Once the review of a new artifact commences, it has to be meticulously analyzed by the administrator. The vetting process at a minimum is as follows:

1. It is required that the artifact follows our standard naming convention. It should be a unique name, containing the artifact type, operating system version, and name of the artifact that is being uploaded. For example: *File Android 7.1.2 Kik 11.40.0 Logs.*
2. It is of highest importance that contributors fill out as many of the fields as possible on the artifact form to create the most complete profile for each artifact. Required fields vary depending on the type of artifact, however, common fields exist among all artifacts, including, artifact and device type, description, search tags, Bibtex (if the artifact was referenced from another source), etc. Nevertheless, artifacts may be submitted with some missing information and an admin member will decide if the profile is comprehensive enough to contribute to the database at that moment.
3. Some artifacts will also have files attached to them, such as text files, plists, html files, SQLite databases, and others. The administrator must meticulously scrutinize each file to ensure that at least three important things are met:
   (a) The information matches the profile.
   (b) The artifact file is sanitized and no Personally Identifiable Information (PII) or PRIVATE material is included, as per the user policies. Sometimes files may contain data that is either encrypted or encoded, such as in Base 64. Finding this information might include decoding any data to plain text to ensure no private information is exposed.
   (c) The most important fields in the file are tagged for quick access.
4. Once the artifact has been thoroughly vetted, the admin will either flag or approve the submitted artifact.
5. If the artifact has been flagged, the user will be notified via e-mail. The e-mail will include any admin notes that were placed on the artifact, as well as things that need to be verified and corrected by the user.
6. Once the user has verified and modified the necessary artifact information, the admin will either approve or continue to flag such artifact until all conditions are met.
7. If, initially, the artifact does not require any revision by the user, then it will be approved and made available for the AGP community to view and search for.

4. Impact

4.1. Professional

As an archival database, AGP serves the digital forensics community by storing and making accessible various types of digital

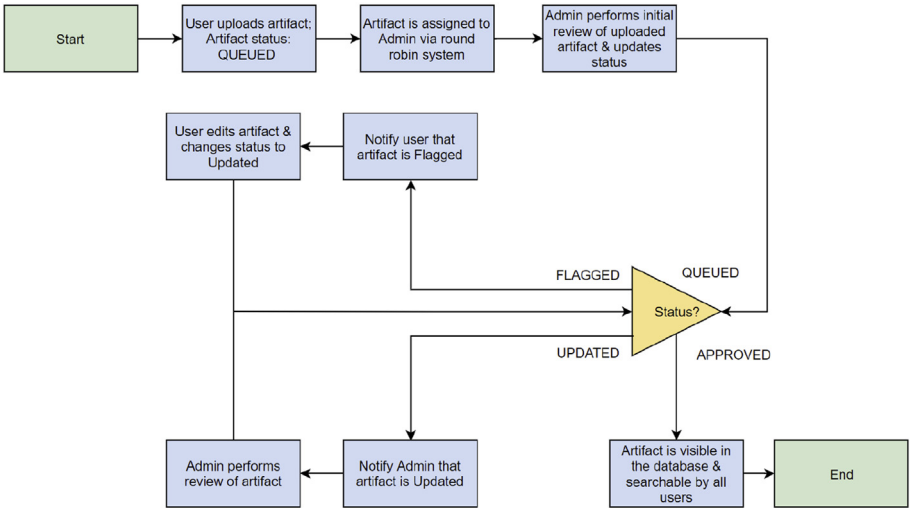


Fig. 5. AGP artifact vetting process.

artifacts. AGP also serves to encourage cooperation in the community. Similar to archeology, digital forensics relies on the research and work of its stakeholders to understand human activity by way of recovered artifacts. Unlike archeology, however, where artifacts and the knowledge obtained from them are shared with others after conducting significant research, such as with the public in a museum, digital artifacts located and studied by practitioners are often kept private.

While practitioners may publish the results of an artifact-centered study, often the artifact itself is not made available to others outside of the study. Work by Grajeda et al. (2017) showed that in digital forensics, less than 4% of researchers who create their own dataset will share it. As the author explains, reasons for not sharing vary, and include not having the resources to do so, not recognizing the importance of sharing datasets, and privacy and propriety concerns.

The lack of dissemination of information and artifacts among the digital forensics community can be an impediment, stunting knowledge building and utilization. As Grajeda et al. (2017) explains, it creates a disadvantage leading to "low reproducibility, comparability, and peer validated research". Without releasing artifacts to the community, practitioners cannot verify results obtained by others or compare their own results to those of others. This could lead to practitioners having to reinvent the wheel, creating their own datasets, which may stall cases as investigators attempt to understand the artifacts they are encountering. Furthermore, practitioners may not be able to keep up-to-speed with the constant release of new devices and applications - a necessity in the field.

As the sharing of knowledge is beneficial for all, AGP intends to increase cooperation within the digital forensics community. AGP has implemented three mechanisms for accomplishing this: friendly competition, tagging, and communication. Acknowledging that competition is a central element in the sciences, AGP has established a leaderboard displaying the number of contributions made by organizations and specific users. For every submitted and vetted artifact, a point is given to the contributing user and their respective organization. Visible only to registered AGP users, the leaderboard is intended to foster friendly competition, thus motivating users and their organizations to freely and willingly share their work with others.

Artifact tagging allows users to provide keywords and brief descriptions when curating an artifact. This allows other users to search for and locate artifacts in the system. Additionally, a user can tag other users, thus providing recognition for those who have collaborated on the research or for those who have done similar or related research.

AGP has also employed messaging via the system to provide users with an effortless mechanism to communicate with each other. Such a platform presents users with the opportunity to seek/offer assistance, provide/receive feedback, and ask/answer research related questions. Essentially, the messaging system is a tool to promote dialogue among the digital forensics community. There is also the potential for networking and recruiting via AGP messaging. For example, organizations involved with AGP and seeking to hire qualified candidates, especially students, may reach out to a user who has actively contributed to AGP.

In addition to a messaging system, AGP distributes a monthly newsletter to communicate with all registered users. Included in the newsletter are details regarding newly uploaded artifacts, the top artifact contributors, the top artifact searches, the latest publications, and artifacts of interest. Calls for artifacts are also often made in the newsletter to encourage submissions from users.

4.2. Academic

One of AGP's objectives is not only to be employed by practitioners and investigators, but also by students, who have been the main contributors of artifacts to the platform. As we have previously mentioned, AGP is a community effort where all participants are able to upload and download approved artifacts. However, as we have analyzed in the months since the launch, AGP has been increasingly beneficial to our own students and collaborators from other universities.

AGP's artifact inventory has grown because it has been rooted in academia. Rather than relying on all stakeholders to share artifacts so that AGP can be sustained, we see future students as the driving force behind these contributions. For example, in Fall 2017 the University of New Haven (UNH) partnered with the University of Texas at San Antonio (UTSA) under a mini National Science Foundation (NSF) grant, which permitted students to earn a salary while rendering their services as artifact diggers.

The student artifact diggers not only conducted their own research to discover, sanitize and upload new artifacts, but they also gained hands-on knowledge and experience in the process.

In Fall (2017) AGP was also implemented in one of UNH's courses. This approach helped AGP surpass the 1000 artifact mark. Some of the new artifacts originated from a one-time class session in Dr. Ibrahim Baggili's Small Scale Digital Device Forensics course. Here, all the students participated in a lab to forensically dig artifacts from different devices and applications. Other artifacts stemmed from a small group in the class who chose the task of digging and uploading 500 digital artifacts for their final project. This hands-on opportunity provided students a chance to learn more profoundly about artifact curation and analysis. The artifacts were extracted from several types of small scale digital devices, operating systems, filesystems and software that would have all been difficult to cover in just one course. Consequently, students gained a better understanding of digital forensic artifacts and were able to explore the volume, variety and velocity of digital forensic evidence found in artifacts. For a brief overview of some of their findings see Section 5.

5. Sample of curated artifacts

At the time of writing, AGP held 784 approved artifacts and over 1000 artifacts tagged with different statuses (See Section 3.4.2 for statuses). To illustrate some of the diverse types of archived artifacts in AGP, in this section we cluster artifacts together. Subsections contain artifacts that were either directly created and used in published research or artifacts that were strictly created by the students mentioned in Section 4.2. These artifacts showcase the diversity of devices and applications hosted by AGP.

5.1. Artifacts from interesting devices

One of AGP's objectives is to host artifacts acquired from trending technologies. As Internet of Things (IoT) devices, and others, continue to appear and evolve, it is of benefit to the forensics community to have prompt, well researched information about these types of devices and the type of evidence that can be found on them.

For instance, AGP includes artifacts acquired from a primary study on the forensic investigation of the DJI Phantom III Drone (Clark et al., 2017). Significant evidence was found on an internally mounted SD card, as well as artifacts found through the DJI Go application managing the drone on the mobile device controlling it.

Some of the artifacts, such as the DAT files containing the flight data, were found to be encrypted, however, with the new open-source tool DRone Open source Parser (DROP) developed by the team, they were able to decrypt the files and confirm the locations the drone had flown in. With this and other artifacts, an investigator has the ability to prove drone ownership, the drones whereabouts, and even the users Wi-Fi credentials.

Another timely device in AGP is the Echo Dot 1.0 with Alexa Cloud. The artifacts related to this device where discovered by Chung et al. (2017) during their study. Relevant artifacts include customer information (i.e., Customer ID, name and e-mail), voice history, calendar, card list (conversations between users and Alexa) and others. The artifacts highlighted are of great interest in criminal cases, such as the recent case in Arkansas.[13] Other notable devices with corresponding artifacts in AGP include Smart TVs, Smart Watches, SteamVR (Virtual Reality Artifacts) and Tablets.

5.2. Timely mobile application artifacts

Nowadays, there are millions of applications available in leading application stores, such as Google Play[14] and the Apple App Store[15] (Statista, 2017). These applications include social media, e-mail, messaging, banking, gaming, music, etc. Applications are now inseparable from one's life. One can also imagine all the potential evidentiary artifacts stored by such applications. Applications know and record so much information about users, making it easier for investigators to learn the user's activities, locations, etc. In a case this data could potentially result in a suspect being found guilty or innocent.

Notable applications present in AGP include KeepSafe, Keeper, AppLock, PhotoLocker and others. These applications are designed to hide sensitive information and sometimes they are even used to conceal child pornography. Substantial research was conducted on these applications by Zhang et al. (2017b) and all artifacts found (e.g., passwords, stored media locations, email, etc.) were uploaded to AGP. Other applications include GPS tracking applications such as Pokemon Go, which stores user locations. An interesting finding from a student was that the McDonald's application not only stores a user's personal information, but also records every McDonald's location the user has visited. Furthermore, messaging and social media applications which are often used in child exploitation and trafficking, including KIK, Snapchat, Tango, WhatsApp, and Facebook have relevant artifacts stored in AGP.

6. Data use and analysis

Data related to the use of AGP are collected. Among these are the number of users, organizations, and countries involved, and the number and type of system interactions. At the time of this writing, there were 174 vetted users from 141 organizations and 17 countries (See Table A.2). The system has also experienced 10,809 interactions.

In addition to the above, quantitative and qualitative data regarding the artifacts is also collected. The quantitative data allows us to know how many artifacts are archived. Currently, the count has reached 1000, a milestone for the project. This data is also used to tally participation by users and organizations, which is shared with the AGP community, fostering a competitive environment. The qualitative data provides details regarding the types of artifacts being shared, such as a Facebook Activity file from an Android phone. This indicates to the AGP team what is trending in terms of research and investigative interests.

Lastly, the AGP team also collects and analyzes data pertaining to searches performed by users. While the search function provides professionals with an efficient manner to locate a specific artifact of interest, it also provides insight regarding the investigative needs in the digital forensics community (See Fig. 6 for a word cloud showing search queries by AGP users). When a search is performed, the query is stored. That data is then reviewed by the AGP team for the purpose of anticipating the type of artifacts that are most needed by the community, and to help set an artifact research agenda. With this information, the AGP team can work to ensure that certain artifacts are researched and made available, generally by encouraging other users to upload the artifact, or having the AGP team dig for them.

7. Lessons learned

The AGP platform was launched in June 2017. At the time of this writing, the system has been online for seven months and has garnered attention from media outlets [16–18] and digital forensic practitioners worldwide. The community views AGP as an important resource for investigative processes and learning methods. Based on our experience in gathering information from research and survey results, creating a standard artifact definition, constructing an ontological model, and implementing it all in AGP, we reflect on the things that we have learned below:

- Some digital forensics practitioners can be hesitant in sharing artifacts. Perhaps they are bound by the restrictions of their employer or simply do not have the time to share.
- Digital forensics practitioners, researchers and educators all want access to a curated artifact platform; the demand is high.
- The digital forensics community is fractured in its approach to sharing.
- Academia is a good place for curating digital forensics artifacts.
- Students learn a lot from becoming artifact diggers and it provides them with experiential learning that would be difficult to achieve in a traditional classroom setting.
- Students tend to dig and curate artifacts related to systems and applications they use.

Nevertheless, we face the challenge of most users being more interested in viewing and downloading the data than actually contributing to the database with any new artifacts. As we know, reasons for not sharing artifacts involve privacy and proprietary concerns, which can possibly be the main reasons why users opt to just taking artifacts, but not contributing. Currently, even though the leaderboard shows that universities are the biggest artifact contributors, they only represent 9 out of 134 of the organizations in AGP. Our main objective is to accelerate the involvement of more academic institutions in the artifact

[13] http://www.cnn.com/2017/11/30/us/amazon-echo-arkansas-murder-case-dismissed/index.html. Last Accessed: 1/20/2018.
[14] https://play.google.com/store.
[15] https://www.apple.com/ios/app-store/.
[16] https://www.forensicmag.com/news/2017/06/university-new-haven-launches-artifact-genome-project-digital-forensics-worldwide.
[17] http://wtnh.com/2017/08/21/university-of-new-haven-launches-artifact-genome-project/.
[18] http://www.nhregister.com/connecticut/article/UNH-in-West-Haven-looks-to-revolutionize-11308667.php.

Fig. 6. AGP searched keywords word cloud.

curation process. Our partnership with UTSA in the Fall 2017, is mainly responsible for helping us realize that working with other academia does indeed bring great benefits to all parties involved.

8. Conclusion

Currently in use by 174 users from 141 organizations, our work has illustrated demand for an artifact curation and sharing platform. AGP is an online crowd-sourced repository for archiving and accessing digital forensic artifacts. Our work built the foundations of what digital forensics artifacts are and constructed a systematic approach for curating them.

9. Future work

As AGP continues to grow, it is essential to continue improving it, not only for the users, but also for the staff that manage it. For instance, the subscription and vetting process of users as previously mentioned (See Section 3.4), will become streamlined. Future iterations will reduce it to a one-step process where users will not receive a second e-mail to fill out the vetting form. These primary and secondary user application forms will be merged into a single form, saving maintenance time.

Furthermore, the administrator will now have the ability to consolidate users that have opted to receive our monthly newsletter and extract all user e-mails in a CSV format file. None of these functionalities are available now, making the administrative tasks a little tedious.

We will also implement a “Like” button for artifacts, similar to social media, which will allow users to distinguish popular artifacts. Additionally, the most “Liked” artifacts will be displayed on AGP's Home Page. Furthermore, on the Home Page, trending keywords will be displayed similar to Fig. 6.

Given that the top AGP contributors stem from academia we will focus our efforts on that user base. We will focus on tackling the challenge of integrating artifacts into academic digital forensic programs. Our intention is to do this by creating educational modules with scalable and self-paced exercises. These exercises will be created using current and future artifacts to teach students about their nature and how to locate, collect, and analyze them, thus, better preparing students for real-world investigations. This approach will employ research-based best practices, such as *Automatic Assessment* (AA) (Malmi et al., 2002), *Self-Paced Learning* (SPL) (Tullis and Benjamin, 2011), *Challenge-Based Learning* (CBL) (Cheung et al., 2011), *Inquiry-Based Learning* (IBL) (Lim, 2004; Kim and Yao, 2010; Woolf et al., 2002), and utilizing realistic data (Woods et al., 2011). Realistic data is important for digital forensics and cybersecurity education so that students get exposed to complexities they would face in the real world upon graduation. All current and future AGP artifacts (CuFAs) are, by definition, of potential forensic value to investigations, making them realistic.

Acknowledgements

This material is based upon work supported by the U.S. Department of Homeland Security under Award Number 2009-ST-061-CCI001-05 and the National Science Foundation under Grant No. 1565560. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Department of Homeland Security. We would also like to acknowledge Dr. Nicole Lang Beebe and Shalabh Saini from the University of Texas at San Antonio for their support and contributions. Finally, we would like to acknowledge Sergeant Corey Davis from the CT Center for Digital Investigations (CDI), for his support and contributions to the AGP project. Lastly we would like to acknowledge all the University *of* New Haven students that have been the major contributors of artifacts to the AGP.

Appendix A. Vetted AGP Users

Table A.2: Vetted AGP users by Country and Organization Type.

Type	Academia	Federal	FFRDC	Local LE	Private	State LE	Total
Country							
Belgium		1					1
Brazil	1						1
Canada		1		2	4		7
Cayman Islands					1		1
Finland		1					1
France		1			1		2
India					1		1
Ireland	1						1
Israel					1		1
Netherlands					1		1
New Zealand	1						1
Norway				1			1
South Africa					2		2
Spain	1				1		2
Switzerland		1					1
United Kingdom	2			3	2	1	8
United States	8	7	1	35	42	16	109
Σ Sum	14	12	1	41	56	17	141

Appendix B. AGP Screenshots

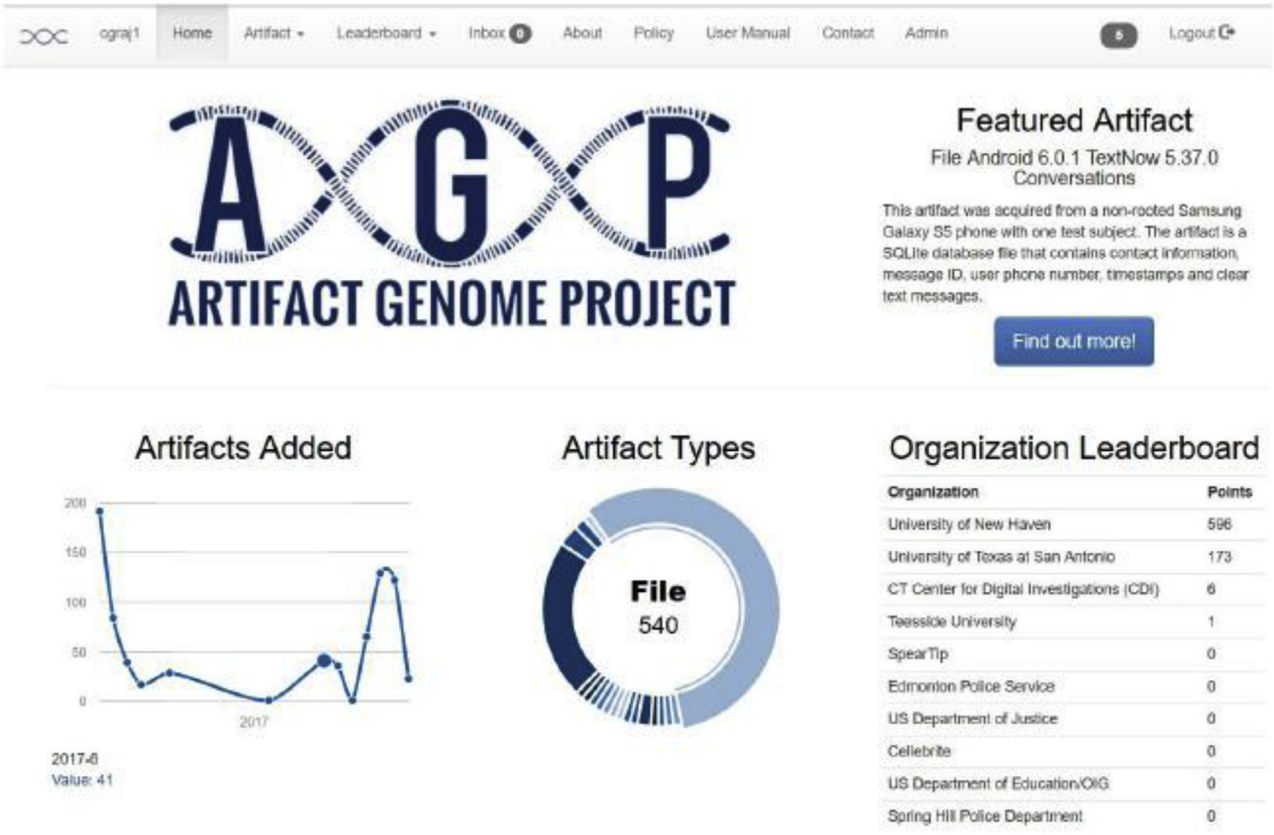


Fig. B7. AGP Home Page, Featured Artifact & Organization Leaderboard

Fig. B8. Sample part of AGP's File Artifact Input Form

References

Ahmed, I., Obermeier, S., Sudhakaran, S., Roussev, V., 2017. Programmable logic controller forensics. IEEE Secur. Priv. 15 (6), 18–24.

Al Marzougy, M., Baggili, I., Marrington, A., 2012. Blackberry playbook backup forensic analysis. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 239–252.

Al Mutawa, N., Al Awadhi, I., Baggili, I., Marrington, A., 2011. Forensic artifacts of facebook's instant messaging service. In: Internet Technology and Secured Transactions (ICITST), 2011 International Conference for. IEEE, pp. 771–776.

Al Mutawa, N., Baggili, I., Marrington, A., 2012. Forensic analysis of social networking applications on mobile devices. Digit. Invest. 9, S24–S33.

Bader, M., Baggili, I., 2010. Iphone 3gs Forensics: Logical Analysis Using Apple Itunes Backup Utility.

Baggili, I., Oduro, J., Anthony, K., Breitinger, F., McGee, G., 2015. Watch what you wear: preliminary forensic analysis of smart watches. In: Availability, Reliability and Security (ARES), 2015 10th International Conference on. IEEE, pp. 303–311.

Bhoedjang, R.A., van Ballegooij, A.R., van Beek, H.M., van Schie, J.C., Dillema, F.W., van Baar, R.B., Ouwendijk, F.A., Streppel, M., 2012. Engineering an online computer forensic service. Digit. Invest. 9 (2), 96–108.

Casey, E., Back, G., Barnum, S., 2015. Leveraging cybox to standardize representation and exchange of digital forensic information. Digit. Invest. 12, S102–S110.

Cheung, R.S., Cohen, J.P., Lo, H.Z., Elia, F., 2011. Challenge based learning in cybersecurity education. In: Proceedings of the 2011 International Conference on Security & Management, vol. 1.

Chung, H., Park, J., Lee, S., 2017. Digital forensic approaches for amazon alexa ecosystem. Digit. Invest. 22, S15–S25.

Clark, D.R., Meffert, C., Baggili, I., Breitinger, F., 2017. 'Drop (drone open source parser) your drone: forensic analysis of the dji phantom iii'. Digit. Invest. 22, S3–S14.

Denton, G., Karpisek, F., Breitinger, F., Baggili, I., 2017. Leveraging the srtp protocol for over-the-network memory acquisition of a ge fanuc series 90-30. Digit. Invest. 22, S26–S38.

Forensic Artifacts (2012). http://forensicartifacts.com/

Garfinkel, S.L., 2013. Digital media triage with bulk data analysis and bulk_extractor. Comput. Secur. 32, 56–72.

Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics–and what is missing. Digit. Invest. 22, S94–S105.

Hale, J.S., 2013. Amazon cloud drive forensic analysis. Digit. Invest. 10 (3), 259–265.

Harichandran, V.S., Walnycky, D., Baggili, I., Breitinger, F., 2016. Cufa: a more formal definition for digital forensic artifacts. Digit. Invest. 18, S125–S137.

Iqbal, A., Alobaidli, H., Marrington, A., Baggili, I., 2013. Amazon kindle fire hd forensics. In: International Conference on Digital Forensics and Cyber Crime. Springer, pp. 39–50.

Kim, D.W., Yao, J., 2010. A web-based learning support system for inquiry-based learning. In: Web-based Support Systems. Springer, pp. 125–143.

Lim, B.-R., 2004. Challenges and issues in designing inquiry on the web. Br. J. Educ. Technol. 35 (5), 627–643.

MagnetForensics, 2017. Artifact Exchange Opens Today for Sharing Custom Artifacts. https://www.magnetforensics.com/blog/artifact-exchange-now-open/.

Malmi, L., Korhonen, A., Saikkonen, R., 2002. Experiences in automatic assessment on mass courses and issues for designing virtual courses. ACM SIGCSE Bulletin 34 (3), 55–59.

Marrington, A., Baggili, I., Al Ismail, T., Al Kaf, A., 2012. Portable web browser forensics: a forensic examination of the privacy benefits of portable web browsers. In: Computer Systems and Industrial Informatics (ICCSII), 2012 International Conference on. IEEE, pp. 1–6.

MITRE, 2014. About Cyber Observable Expression. https://cybox.mitre.org/about/.

NHGRI, 2016. An Overview of the Human Genome Project. https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/.

Quick, D., Choo, K.-K.R., 2014. Google drive: forensic analysis of data remnants. J. Netw. Comput. Appl. 40, 179–193.

Ricci, J., Baggili, I., Breitinger, F., 2016. 'Watch what You Wear: Smartwatches and', Managing Security Issues and the Hidden Dangers of Wearable Technologies, p. 47.

Roussev, V., Barreto, A., Ahmed, I., 2016. Forensic Acquisition of Cloud Drives arXiv preprint arXiv:1603.06542.

Roussev, V., McCulley, S., 2016. Forensic analysis of cloud-native artifacts. Digit. Invest. 16, S104–S113.

Senthivel, S., Ahmed, I., Roussev, V., 2017. Scada network forensics of the pccc protocol. Digit. Invest. 22, S57–S65.

Statista, 2017. Number of Apps Available in Leading App Stores as of March 2017. https://www.statista.com/statistics/276623/number-of-apps-available-in-leading-app-stores/.

SWGDE, 2015. Swgde Glossary. https://www.swgde.org/documents/CurrentDocuments/2015-05-27 SWGDE-SWGITGlossaryv2.8.

Tullis, J.G., Benjamin, A.S., 2011. On the effectiveness of self-paced learning. J. Mem. Lang. 64 (2), 109–118.

Van Beek, H., van Eijk, E., van Baar, R., Ugen, M., Bodde, J., Siemelink, A., 2015. Digital forensics as a service: game on. Digit. Invest. 15, 20–38.

Walnycky, D., Baggili, I., Marrington, A., Moore, J., Breitinger, F., 2015. Network and device forensic analysis of android social-messaging applications. Digit. Invest. 14, S77–S84.

Woods, K., Lee, C.A., Garfinkel, S., Dittrich, D., Russell, A., Kearton, K., 2011. Creating realistic corpora for security and forensic education. In: Proceedings of the Conference on Digital Forensics, Security and Law', Association of Digital Forensics, Security and Law, p. 123.

Woolf, B.P., Reid, J., Stillings, N., Bruno, M., Murray, D., Reese, P., Peterfreund, A., Rath, K., 2002. A general platform for inquiry learning. International Conference on Intelligent Tutoring Systems. Springer, pp. 681–697.

Zhang, X., Baggili, I., Breitinger, F., 2017a. Breaking into the vault: privacy, security and forensic analysis of android vault applications. Comput. Secur. 70, 516–531.

Zhang, X., Baggili, I., Breitinger, F., 2017b. Breaking into the vault: privacy, security and forensic analysis of android vault applications. Comput. Secur. 70, 516–531.