The Internet of All Things: Collecting the Right Data For Your Case

Warren Kruse

Paul McVoy

Kevin Chang

Copyright 2017, The Sedona Conference.
All rights reserved.

The Internet of All Things:
Collecting the right data for your case

Paul H. McVoy, Meta-e Discovery
Warren G. Kruse II, Altep Inc.
Kevin Chang, Meta-e Discovery

Similar to the data deluge of the early 2000s, the eDiscovery world is being confronted with a new challenge as we grapple with what to do with data being created by everyday things. Smaller, faster, cheaper computer chips make it easier to transform almost anything into a data processing and information storage repository, from our cell phones (which now seems like a passé application), to our appliances, watches, cars, toys, medical devices, and our artificial intelligence mimicking, personal home hubs. Various pieces of data are recorded and transmitted to a central server, many times without a person's awareness or knowing consent. The International Data Corporation (www.idc.com) estimates that by 2020 there will be 212 billion connected devices, each with the ability to record and store data. The subset of these new "smart" devices is commonly referred to as The Internet of Things or IoT.

As we are all forced to ride the edge of this tidal wave of disparate data sources, it is crucial to take a realistic view of each type of potential responsive information repository and know what is available, how easy or difficult it is to get to, and what meaningful value does that data have to the case at hand. Only by understanding that while my internet connected refrigerator records and tracks its contents, and how often I let my milk run out, it likely has nothing to do with my company's participation in a class action related to price fixing.

Devices and Data

Smart Phones. Smart phones were the first major source of alternative, potentially responsive electronically stored information (ESI). The industry struggled with what data was on the phones and then with finding tools to accurately get the data from them. Adding to the issue were the various phone operating systems, each of which was highly proprietary and closely guarded by their developer. Setting aside the troubling issue of accessing locked phones and the developers' objections to help, many of the technical hurdles have been overcome with regard to data acquisition, which now enables us to assess what kinds of data are available so we can determine our need to preserve and collect it.

> Email: Smart phones allow for the management of email. Most do this by syncing with an email server. Often a single phone will be linked to several different service providers, to manage a person's work and private email. The good thing about email is that most times, it will reside in a location other than on the phone itself. Sent and received messages are usually synced to a server that hosts the email service; a work enterprise server, a public cloud or to a user's own personal computer. While it is important to confirm this when assessing the usage of a phone that may have potentially responsive ESI, this is usually a non-issue. The one area that will sometimes require special handling is when the phone is the only source of a specific email, like a draft.
>
> Instant Messages. Text messages have quickly become the preferred method of communication for many people, both for their personal and professional interactions. Unlike email, oftentimes the phone, tablet or other mobile device is the only source for these message chains. It is important to quickly assess the potential responsiveness of text messages so that they can be

Copyright 2017, The Sedona Conference, Warren Kruse, Paul McVoy, and Kevin Chang

adequately preserved and collected in a usable format. Historically, text messages were difficult to preserve and collect, which led to wholesale exemptions from the discovery work flow. This is no longer the case, and should be considered as potentially discoverable ESI.

Application or App Data. Various applications can be installed onto a smart phone to do everything from assisting with business tasks to playing games. Each of these applications leave some data resident on the phone. Most times, this data is not a valuable source of information in litigation, but it can't be discounted entirely. A document drafting app may be the only source of a document draft. Some social media apps, like Facebook, even have their own communication and messaging functionality. While this data can sometimes be retrieved from the application providers' servers, in some instances this data will only appear on the phone. Also, consider that some apps purport to erase any data after only a few seconds or minutes, but that may not always be the case or the user may have downloaded another app to store the otherwise ephemeral information.

System Data. Smart phones track a number of things "behind the scenes." For example, smart phones can record every cell phone tower and Wi-Fi hotspot you encounter and connect to. Pictures taken on a smart phone can have embedded geolocation and temporal information inserted into the picture's metadata. System information is another aspect of new data that needs to be considered when assessing ESI for a specific case. Picture metadata has often been found to be valuable sources of information when trying to connect a person to a location.

Wearables. Activity trackers and smart watches are another source of potentially responsive data that need to be considered for preservation and collection. Wearables often have storage built in, so even though much of the data is also kept on the cloud or your phone, each device can have gigabytes of storage internally. Like smart phones, data can be comprised of text messages and email but can also include health information, such as heart rates over time and sleep patterns. Watches can also download applications and store app data as well.

Appliances. Home appliances are now often connected to the web. Televisions were among the first to become "smart" but now people's refrigerators, lights and thermostats are regularly part of the Internet of Things. And each of these devices can collect and store many types of information, from your viewing preferences to your activity in and around your house. For the most part, many of the devices only store account information that it needs to perform the purpose for which it was designed. For example, a television might be connected to the internet and allow you to stream movies or access the internet, but generally the only thing being stored is the log in credentials to those services.

Cars. Cars have long had sophisticated computers built in, and as part of those systems, there have been data storage used to track speed and various performance parameters of the engine. Data from these systems has been used in court before, to prove either the improper handling of the vehicle or that the vehicle was located in a specific place that it either should or should not have been.

In addition to GPS information, cars can also maintain several "black box" like components that survey various systems in the car, reporting and recording performance metrics that are intended to be used for maintenance and repair but that could also be used as evidence in products liability cases. One such example is the litigation regarding electronic components manufactured by Bosch and singled out as a problem in the Volkswagen emission cases. As reported in the Financial Times on October 5, 2016 "The allegations against Bosch focus on its electronic diesel control unit 17, a component supplied to VW and

capable of gathering data on vehicle speed, acceleration, air pressure and the position of the steering wheel." Lawyers in the case alleged that this device was programmed to manipulate output when the device recognized it was being tested. (https://goo.gl/Yk7kvj)

Virtual Assistants. The cutting-edge trend for IoT devices is the virtual assistant. The Amazon Echo and the Google Home are the most well known. These devices work by passively listening for a "wake" word, and upon hearing the word record your voice and translate that into a command that it stores internally and on the service provider's servers. It is unclear how much of the passive time is also recorded, but it is relatively simple to access the recorded commands that you have given it. Data that is stored on one such device has been part of an ongoing murder investigation in Alabama. According to an article on the Verge website, police were able to extract "audio recordings, transcribed records, text records, and "other data"" from an Echo. (https://goo.gl/8AqJGv). We will likely see similar demands for data recoreded by these devices.

Home Hubs. As users add more and more IoT devices to their homes, companies are developing connected home systems managed by what are being called hubs. While individual appliances may not store much data, and rely on storage to be managed on the cloud, this new class of devices is designed specifically to connect to your personal IoT and maintain that connection even if your internet connection is lost. Practically, this means that the hubs have storage contained in them. Data that can be found on these hubs are information about your connected devices (the times your lights turn on and off, the day and night temperature settings of your thermostat) as well as other information that could be potentially discoverable. For example, some home hub systems have as their centerpiece your home's security system, including storage for surveillance cameras.

Preservation and Collection

After assessing and understanding what data is potentially available, there are the practical issues dealing with getting the data out of the devices in a defensible and reviewable manner.

In the past, the burden of preserving potentially responsive ESI has typically been comprised of the cost to maintain large repositories of existing and archival data. A new challenge is presented with the growth of the IoT. Oftentimes backing up and retaining data from these devices is a technologically complex endeavor requiring the expenditure of a disproportionate amount of financial and human resources. In addition, these devices and their associated networks are frequently not designed for long term data storage and retrieval, adding to the costs by requiring new methods of preservation and collection to be developed for each new device.

Furthermore, with the increasing incidents of cyber-crime, like hacking, data breaches and identity theft, companies are making it harder and harder to preserve and extract data from the devices. For these reasons, it is highly advisable to first make a realistic assessment of the data that is necessary for your specific case and then to work transparently with the opposing counsel to develop a proportional and realistic work plan.

Generally, IoT connected devices store potentially responsive user data and evidence within databases that are located on the device. The most common database format used for this purpose is SQLite. SQLite is an open source, server-less transactional database engine that allows high customization for little cost, and with no limits on its use, be it private or commercial.

(https://www.sqlite.org/about.html). These types of databases are attractive because they can be completely self-contained and do not require access to a server for processing commands or managing the storage of data.

Mobile forensic tools, as well as standalone database forensic software, are used to parse the contents of these databases for reporting, meaning that it is possible to target specific types of data for collection from certain devices. It is also important to understand that these databases may also contain deleted records that have not yet been purged. The user of the device might not even be aware that data they had thought was deleted is still on their device and susceptible to preservation and collection.

As with traditional data sources, the acquisition of IoT devices can be performed using different collection methods, each having its own benefits and costs. When determining which method to use, it is important to factor what the data will be ultimately used for, the volatility of the source data, and the needs to the case. Cost can also become a factor and needs to be weighed.

> Logical Collection. Active data is collected from the device, which means that no forensic processes are undertaken to discover deleted data. Deleted data may still exist on the apparatus's database, as described above, but will not be collected during a routine logical collection.
>
> File System Collection. This method is similar to a logical collection, but the data collected will also contain all contents about the file system used by the device, the operating system and the databases.

Physical Collection. This is the most complete, and therefore complex and resource intensive method of collection. In this process, both active data is collected as well as deleted data that may still reside on the device. Also, the device's file system is captured in its entirety, including any slack space that may be empty but may also have once stored data.

It is important to keep in mind that, similar to the function of a computer, when a user deletes a file or piece of data, only the pointer to the file is deleted. The storage location of the orphaned data is then made available to the system's operating system for new data. If nothing is copied over the original data, it is possible for that data to be identified, analyzed, and made part of a discovery process.

The collection of data from IoT connected gear is further complicated by the variety of devices and the security measures the various manufacturers place on them to protect their clients' privacy and data. Add to this challenge that new gadgets are frequently introduced to the market that collection professionals have never encountered before. Sometimes, the only way to capture potentially relevant information from these sources is to make a video recording of a technician scrolling through the data so at least the substance of the information can be captured and preserved.

Another potential source of data from IoT connected items are the mobile device's backups to which the item is connected. As discussed above, oftentimes the storage for the IoT happens on a smart phone or in the cloud. When the phone is backed up, it can create a snapshot in time of the data the specific device was storing. Like an email backup system, data can be recovered from a time in the past that may no longer exist, or from a device that no longer has a copy of the data. Also, because the tools and processes for accessing these types of backup data already exist, it is sometimes easier to collect potentially responsive information from a backup than it is from the device itself.

Legal Considerations

The law with regard to the Internet of Things is in its early stages of development. Aside from the few blockbuster cases involving data on locked devices sought by the Federal Government or the sensational murder case referenced above, there have not been many published opinions regarding the preservation and collection of data from the IoT. However, there is some helpful guidance out there and some lessons learned from the preservation and collection of unique data sources that can be applied.

Admissibility of Cloud/IOT Evidence

As a threshold matter, the collection of digital evidence must satisfy certain expert evidence standards for admissibility. Rule 702 reads: "If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill experience training or education, may testify thereto in the form of an opinion or otherwise." In other words, scientific evidence shall be considered competent if it possesses a basis in the methods and procedure of science. To determine whether this is the case, the Daubert court proposed several illustrative factors:

- Whether the theory or technique employed by the expert is generally accepted in the scientific community;
- Whether it has been subjected to peer review and publication;
- Whether it can be and has been tested;
- Whether the known or potential rate of error is acceptable; and
- Whether the research was conducted independent of the particular litigation or dependent on an intention to provide the proposed testimony. Daubert v. Merrell Dow Pharmaceuticals, Inc. 43 F.3d 1311 (9th Cir. 1995).

Digital forensics is highly technical and relies on multiple scientific disciplines (computer science and computer engineering, along with underlying associated mathematics and physics) as well as the highly-specialized knowledge and judgment of professional information technologists and system engineers. That said, with *Kumho Tire Co. v. Carmichael*, Daubert-criteria has also been extended to non-expert scientific expert evidence. Under Kumho, digital forensics evidence may be tested for admissibility by virtue of expert acceptance of the tools being used, comparison with articulated standards, known error rates, and other factors.

Beyond a Daubert-type challenge, the collection and processing of data may also have to survive a competency inquiry to ensure that the evidence is properly preserved and presented in compliance with the Best Evidence Rule. Rule 1003 on the admissibility of duplicates governs here, as the producing party is presenting not the storage media themselves, but files readable to ordinary persons. If a digital forensics expert performs the collection and chain of custody is properly monitored, this is typically enough to surmount any challenge with regards to accuracy of the produced data.

Problems may arise, however, under both Daubert and the Best Evidence Rule, with regards to the newer forensics challenges of the IoT. The fact is, collection methods and capabilities in both these areas reside in largely uncharted territory: the service providers themselves may not be familiar with the needs of eDiscovery and vendors' processes and experts have not been properly vetted. Moreover, the technology of the collection tools hasn't caught up to the vast variety of usage. IoT data types can be challenging to process or necessitate conversion into an alternate file types for human review and metadata is frequently difficult to retain. As collection processes take the necessary time to further normalize, it's advisable that organizations require that their IoT and cloud storage providers be very upfront of their capabilities well in advance of any anticipated litigation.

It is also advisable to enter into discussions with opposing parties once an assessment of potential IoT devices has been made. All parties will face similar challenges, and transparency will prove to the court that best efforts are being taken should issues arise.

Discovery of ESI Held Overseas

As discussed above, devices on the Internet of Things allow for the deployment of a variety of applications, many of which store their data on the device and others which use the cloud as a primary storage location. There are presently no regulations preventing this data from being stored in countries other than the United States. Consequently, discovery of data in the IoT may lead to data stored internationally, triggering other considerations.

Data stored, in whole or in part, outside of the U.S. is subject to the discovery regulations of the foreign jurisdictions where the data resides. Cloud storage further complicates matters as oftentimes a client may not even know where their data is stored. While the U.S. has a fairly liberal discovery regime that encourages production of information, most of these foreign jurisdictions have far more restrictive rules. The European Community, for example, has stringent regulations regarding how personally identifiable information such as gender, marital status, nationality, and identification numbers may be collected, processed, stored, and disclosed. Several European countries have enacted legislation specifically designed to shield their citizens from U.S.-style discovery. These types of regulations present a distinct challenge to the production of relevant electronically stored information in compliance with the Federal Rules of Civil Procedure.

Organizations must consider a number of legal issues when called upon to produce such information. First and foremost is whether the foreign jurisdiction at issue does in fact have regulations regarding the overseas transfer of electronically stored information. This will naturally depend on the laws of the nation in which the data is located. Also of interest is whether courts will be sympathetic to the challenge of collecting data from foreign jurisdictions with these types of restrictive rules. In general, judges have been somewhat lenient in these situations, giving some deference to localized privacy laws (though they are less likely to give deference to foreign blocking statutes).

An organization can also raise the threshold question of whether ESI stored overseas can be said to be under its "possession, custody, or control" under Rule 34(a). Circuits have split and courts typically apply one of three tests to determine whether a party has control over ESI: the "Legal Right" test, the "Legal Right Plus Notification" standard, and the "Practical Ability" standard. Under the "Legal Right" test, a party is said to have control over information if they have the legal right to obtain it. The "Legal Right Plus Notification" standard follows the same guidance, with the additional obligation to inform the requesting party if a third party is in possession of the data. Finally, under the "Practical Ability"

standard, a party must produce information requested in litigation if it has the practical ability to obtain the documents or ESI whether or not it has the legal right to obtain them. The "Practical Ability" standard is particularly dangerous for organizations with discoverable information overseas as it may compel them to violate foreign data privacy laws.

In order to at least help navigate the above issues, an organization can engage in certain best practices in anticipation of future overseas discovery. They should communicate with their vendors to know where their data resides and familiarize themselves with relevant jurisdictional regulations; engaging local counsel during the collection process is highly recommended. Cooperation with opposing counsel and being upfront about issues may help in trying to limit the scope of discovery and avoid problems. Performing review on-site and making necessary redactions can also minimize the amount of data that must eventually be moved out-of-country. Finally, organizations should be careful to ensure that data is secured once transferred to the U.S. by contracting with vendors who are capable of implementing the necessary protective measures.

BYOD

The issue of "possession, custody, or control" is not limited to cloud data held in foreign jurisdictions but also pertains to personal items used in a workplace capacity. This "Bring Your Own Device" phenomenon most notably occurs with regards to personal cell phones used by employees for business purposes, but now can spread to other Internet of Things devices. It is becoming more and more common for employees to bring their Echo to work to listen to music and perform web searches. As noted above, data is recorded on these devices and may need to be considered in a preservation and collection effort.

In most matters, an organization does not have "control" over the personal gear of its employees nor the legal right to access their personal data held therein. BYOD policies potentially obfuscate the issue, however, where employers can access these devices for work-related information and agree to allow employers to either access the work-related information on their phone, or install applications on the device to wipe any non-personal data. Organizations can also utilize a Mobile Device Management (MDM) system that will inventory employee owned gear attached to the entities data storage or operations systems. If enacted, it can be one way of assessing the exposure of IoT devices that are potentially responsive to a given matter.

In order to maintain adherence to privacy laws, organizations should be careful about what data they're entitled to collect and how they handle it. Strong workplace policies that aim to segregate personal and company data should be implemented and enforced. Collection workflows should aim to minimize the amount of personal data being introduced to the discovery review process. Finally, organizations should be cognizant as to where there may be alternative sources of the same information, such as email stored on company servers. The best option is that collection from personal devices may be avoided altogether.

Summary

While the Internet of Things devices are becoming a prolific aspect of modern society in terms of communication, both active and passive, as well management systems for daily activities once available only on computers, it is critical to keep their utility in any specific matter in perspective. IoT gear has the

ability to store our data, text messages, emails, photos, videos, web browsing histories, and our location, but the determination as to whether that data is potentially relevant, and whether the discovery of that data is proportional to the matter at hand, is vital to the realistic preservation and collection of IoT data.

The first step is to assess the devices in question as it specifically relates to the case at hand. If, after careful consideration, you determine there is potential ESI residing on IoT devices, it is best to open a dialogue with your opposing counsel to determine how and what data from these devices will be handled.

Guidance can be gathered from the Sedona Conference Database Principles, which advocate for an open dialogue on the preservation and production of structured data. The discussions are encouraged to be transparent because failing to do so can create issues regarding wasted effort and time spent on what may otherwise be a non-issue had the parties come to agreement early. Data contained in the Internet of Things should be handled similarly, as much of it resides in database and in proprietary formats that need to be normalized before being able to be reviewed and produced. The cost to do these things is potentially significant and unless both parties agree it is necessary, by fully knowing the true barriers and pitfalls, informed decisions cannot be made. Ultimately, it may be for the courts to decide, and if so, being able to articulate what you may have, and the difficulties in getting it, will be necessary as well.