[*] starling lab

Best Practices for Web Archiving

White Paper - Nov 2022

Scott Martin, JD
Global Justice Advisors

Basile Simon
Starling Lab, Stanford EE

Digital web and social media content have emerged in recent years as some of the most critical evidence that investigators and prosecutors use in domestic and international courtrooms throughout the world. Particularly in the case of international criminal tribunals, such evidence has been increasingly useful in establishing legal responsibility for the commission of international crimes,[1] for two reasons:

- New investigative and reporting techniques have emerged and are now widely used by civilian investigators[2] and prosecutors[3] alike ;
- Both the aforementioned techniques and recent conflicts such as Ukraine and Syria rely heavily on social media and social messaging as means to documenting and collecting evidence.

Capturing web content as a research resource or for probative means is now commonplace, and it is now incumbent upon investigators, prosecutors, defense counsel, judges, and others to improve their understanding of digital evidence – as well as of the risks associated with it, particularly regarding its authenticity and provenance. This includes best practices for identification, collection, verification, storage, and preservation so it can be used in legal accountability proceedings throughout the world. This Whitepaper seeks to contribute to this discourse by addressing challenges that exist in archiving web pages and presenting best practices.

This Whitepaper will:

- Summarize a technical workshop where participants discussed the **cutting edge techniques and state of the art in web archiving**.
- Summarize a legal roundtable where participants considered existing web archiving practices and **discussed the potential for web archive material to be used in legal proceedings connected with international crimes.**
- Describe the **International Criminal Court's rules of evidence and recent cases addressing the admissibility of digital evidence**.
- Consider **best practices regarding web archiving** and **recommend trial techniques and practices** that endeavor to improve the chances that archived web pages are accorded high probative value.

[1] Daragh Murray, Yvonne McDermott, K Alexa Koenig, Mapping the Use of Open Source Research in UN Human Rights Investigations, Journal of Human Rights Practice, Volume 14, Issue 2, July 2022, pp. 554–581.
[2] The Economist, The people's panopticon: Open source intelligence comes of age, July 7, 2022 issue
[3] Dutch Prosecution Services, The MH17 criminal files

Summary of the Two Workshops

In the fall of 2022, the Starling Labs for Data Integrity ("Starling Labs") held two workshops – one technical and the other legal – that sought to bring these two professions into conversation with one another over how cutting-edge technical advances in web archiving may bolster digital evidence that will be used in accountability proceedings. Together, the technical and legal workshops represent two halves of a joint effort to bridge these professional worlds. Doing so should improve collaboration, mutual understanding, and contribute towards ensuring that critical digital content can be collected, stored, authenticated and later used to advance justice for victims of Russia's war against Ukraine.

On 25 August, Starling Labs convened a group of web archiving experts to discuss practices that could be used to preserve information for accountability purposes in Ukraine. An array of collection, authentication and preservation strategies were presented – with a focus on approaches that best preserve the integrity of recorded web pages (and/or other archived digital material). First, participants explored existing web practices and how they work on a technical level. Then experts weighed on the technical risks to web archives that may endanger the preservation of their integrity. Participants paid particular attention to the risks of storing web archives in existing archival models and how a more distributed, decentralized model may more often than not bring real benefits in terms of long-term resilience and availability.

Having fleshed out the technical foundations of the practice, a roundtable of legal experts – both lawyers specializing in war crimes cases and legal professionals having experience working with digital evidence – convened on 27 September to discuss the legal dimensions to the web archiving practices. Participants sought to identify possible legal vulnerabilities with current archiving practices and to consider the extent to which such material could be expected to be admitted into evidence in a courtroom and given weight at various stages of future criminal proceedings. Participants then articulated best practices that ensure that web archive data are capable of being preserved, produced, and authenticated in a manner that preserves their integrity – and consequently enhance their reliability, utility, and probative value as evidence in a curial context.

Results of the Workshops

This section is an effort to bring the takeaways from each workshop together into: (i) a summary of how the rules of evidence interact with existing web archiving techniques; (ii) a cohesive set of recommendations for how the methodology of web archiving can be designed with an eye towards admissibility; and (iii) a list of techniques that litigators can leverage to ensure that courts admit the digital content into evidence and adequately weigh its probative value.

1- Rules of Evidence and Web Archiving

The rules of evidence (and their interpretations in the jurisprudence) govern whether a court will deem a proffered web archive authentic, admissible, and deserving of proper weight at the decision-making stage of the proceedings. Accordingly, significant time was reserved during the legal roundtable discussing these rules and what they teach us about best practices when seeking admission of web archives into evidence. The conversation was focused on the rules of evidence and practices at the ICC, as this court may prove to be the most consequential in prosecuting high-ranking political, military, and security officials. Participants also discussed specific cases in which digital information (webpages, social media posts, and other information) was accepted into evidence and used to assist the ICC Prosecutor in cases against an accused person. Participants also discussed cases in other jurisdictions, including the United States, Poland, and Germany.

1.1 - At the International Criminal Court

Article 69(4) of the Rome Statute lays out a three-part test governing the admission of all evidence, including any digital evidence or web archive material, for trials conducted before the ICC. It provides that the ICC Trial Chamber may "rule on the admissibility or relevance of any evidence, taking into account, *inter alia*, the probative value of the evidence and any prejudice that such evidence may cause to a fair trial or the fair evaluation of the testimony of a witness".

According to the three-part test, if the evidence is deemed *prima facie* relevant (i.e. it assists in proving the existence of a fact material to the proceedings), then the court assesses its probative value (i.e. part two). This evaluation considers, *inter alia*, the authenticity of the item presented – a critical factor in deciding whether it will be admitted into evidence. If deemed authentic, so long as it does not unduly infringe upon the fair trial rights of the accused (i.e. part three), the item will be admitted into evidence.

Web archivists have particular control over authenticity as it relates to the admissibility inquiry: the choices they make from capture to storage, and the diligence they afford to permit the verification of the material, directly impact the likelihood of admissibility and overall weight of the evidence they collect. As discussed later in the white paper, it is critical to keep this in mind when developing archival frameworks.

On a practical note, evidentiary practice at the ICC is traditionally flexible and permissive regarding the admission of documents into evidence. The admissibility threshold is particularly low in relation to common law courts (such as England and Wales or the US federal courts). This is partly how the practice evolved organically, but also a recognition that civil law systems (such as in continental Europe, where nearly all States are civil law-based) tend to be more flexible than common law systems in this regard.

But the weight that is accorded a particular piece of evidence can be variable. Questions as to the ultimate weight to be given to any given piece of evidence are frequently reserved until the trial chambers' final assessment once all the evidence has been heard. This allows judges a degree of discretion to operate within the framework set forth by the rules of procedure and evidence and avoids an excessively dogmatic approach. This is possibly an allowance for judges - who often come from very different legal systems, cultures, and histories than their colleagues - to continue to rely upon their domestic evidentiary rules and practices while balancing them with the other judges on the bench.

Accordingly, it is likely that in ICC proceedings web archive materials will face fewer successful challenges at the admissibility stage and the critical point of focus will be what is made of such evidence in the Court's final determination in any given case. This prediction is supported, for example, by the case *Prosecutor v. Jean-Pierre Bemba Gombo*, where screenshots of Facebook postings of the Accused's sister were entered into evidence.[4] In *Prosecutor v. Al Hassan Ag Abdoul Aziz Ag Mohamed Ag Mahmoud*, the Trial Chamber admitted a geolocation report that relied upon digital evidence, as it "would be of assistance to the Chamber in its assessment of audio-visual material which allegedly depicts locations in Timbuktu at the relevant time of the charges".[5] In the *Prosecutor v. Ahmad Al Faqi Al Mahdi*, the Prosecution had YouTube videos and online images entered into evidence (this was not even contested by the Defence).[6]

[4] This was despite valid objections made by the Defence, who made three core objections: "First, the Prosecution provides no material supporting the attribution of the Facebook pages to D-6 or P-0274. Since the creation of a Facebook account does not require any valid identity information, it is impossible to forensically ascertain, even on a prima facie basis, that a Facebook account under a certain name is attributable to a person of the same name".

Second, the photographs are not genuine extracts but merely screenshots of a webpage with a pop-up photograph. Unlike a genuine extract, the metadata of the photograph, such as the creation date, the photographing device and the modification traces, are not available, which warrants their exclusion. Such information is particularly relevant since the photographs are used in conjunction with events on a particular date.

Third, the photographs have no probative value at this juncture. The identification of the persons in the images is not yet in evidence. The Prosecution has also failed to provide any explanation or justification as to why this material is not being tendered through a witness".

[5] *Ibid* (other citations omitted).

[6] Rafael Braga da Silva, Updating the Authentication of Digital Evidence in the International Criminal Court, International Criminal Law Review (2021) 1-24, p. 10, *citing Al Mahdi* 22 August 2016 Transcript...p. 29, lines 3–4, pp. 45–46, lines 1–2, p. 48, lines 10–12, and p. 113, lines 6–11. Page 113 provides: "[t]his report, it's entitled "geolocalisation" -- or, "Geolocation of Videos & Images" and that's exactly what it is. It's looking at details on videos and images and comparing them to other images and videos, and directions in order to determine where the place is. I actually, for myself, it was very helpful to reading in the introduction the other description of what geolocation means, which is -- which is described as visual verification".

1.2 - In Continental Europe

Many countries on the European continent are taking a leading role in the investigation of international crimes perpetrated during Russia's war of aggression. This includes the Czech Republic, Estonia, France, Germany, Latvia, Lithuania, Norway, Poland, Spain, Sweden, and Switzerland, who have all initiated criminal investigations concerning crimes committed in the course of the Russian invasion. Investigators in these countries have largely met with Ukrainian refugees to document crimes perpetrated against them since 24 February 2022. Leveraging the information they have collected, investigators can take three principal actions:

- Prepare the information and submit it to the ICC;
- Prepare the information and submit it to the Ukrainian Prosecutor General's Office; or
- Prosecute the actions domestically under universal jurisdiction principles (or, if available, other jurisdictional heads recognised in international and/or domestic law).

These investigations may lead to domestic prosecutions, so it is useful to consider admissibility practices concerning web archives / digital evidence in several representative countries on the continent. In this respect, the Legal Roundtable considered Germany and Poland. Ms. Karolina Aksamitowska, a professor at Tallinn University, noted in a recent article that user-generated digital evidence in Germany is rarely introduced into evidence on its own, but once corroborated by other types of evidence, such material may be accorded significant probative value in German courts. She further notes that "open source evidence [which would include web archives] is often introduced in conjunction with files found on electronic devices and flash drives".[7]

As a civil law country, such permissive evidentiary standards should come as no surprise. But the lenient admissibility rules in some jurisdictions should not be viewed as license to engage in poor web archiving practices like relying on low-fidelity screenshots or not including digital signatures, for example. The objective in web archiving is not merely to clear admissibility thresholds (which are sometimes very low), but to collect and retain evidence of high probative value. In this respect, preserving web archives should always follow best practices (as described later in this document), increasing their prospects of being used effectively to bring accountability and justice for victims of the war in Ukraine.

In Poland, the courts have accepted digital evidence in the courtroom without much hesitation. Generally, there are no formal restrictions on the types of digital data (e.g. text, images, metadata, social media postings, or others) that may be used by parties in legal proceedings as evidence. While substantive technical guidelines on the admissibility of digital evidence may not be present in rules of evidence and procedure, they are described in various international and Polish forensic literature on the subject.

[7] *See* Journal of International Criminal Justice, Digital Evidence in Domestic Core International Crimes Prosecutions: Lessons Learned from Germany, Sweden, Finland and The Netherlands, 16 May 2021.

Ms. Aksamitowska notes that digital evidence was presented in 253 out of 370 legal proceedings she considered while conducting her analysis of digital evidence in Polish courts. Open-source information was the most common type of digital content used in criminal proceedings.[8] Importantly, she notes that "in none of the analyzed cases [was] digital content inadmissible or unreliable for any reason".[9] Types of digital content reviewed included data openly available online (characterized as OSINT by the author); emails; text messages on social media sites; text messages via SMS; and others.[10]

1.3 - In the United States

A discussion of the rules of evidence and evidentiary practices for seeking to admit web archives into evidence in the US Court system also took place during the roundtable. However, even though the United States leads the trans-Atlantic response to investigate Russian war crimes through its leadership on the Atrocity Crimes Advisory Group,[11] US authorities are not playing a significant role in the investigation or adjudication of international crimes in Ukraine. This is likely due to the limited number of Ukrainian refugees in the United States (and therefore fewer individuals to interview), traditional diplomatic task-sharing amongst allies (the US is leading on many initiatives, but Europe leads regarding legal accountability efforts), and the absence of universal jurisdiction in the United States Criminal Code.[12] Nevertheless, jurisprudence and litigation practice from the US Courts often play a role in advancing a global understanding of evolving notions of law and evidence. This is particularly the case with web archives, as hundreds of cases have addressed their admissibility. Accordingly, the roundtable briefly described how web archives are treated in the US federal court system.

Mr. Nicholas Taylor led this discussion,[13] having reviewed admissibility trends in US federal courts, while considering a variety of techniques relied upon to seek the admission of web archives, specifically from the Wayback Machine.[14] In general, web archives from the Wayback Machine have been successfully entered into evidence if accompanied by an affidavit prepared by the Internet Archive.[15] Litigants have also sought to admit such evidence via witnesses with direct personal knowledge, expert witnesses, and/or by way of judicial notice. Judicial notice, which permits the admission into evidence of something so well known or authoritative that its authenticity is not subject to reasonable dispute[16], has been widely used, but its success as a strategy for admission and authentication of

[8] *Ibid.*

[9] *Ibid.*

[10] *Ibid.*

[11] *See* US Department of State, The European Union, the United States, and the United Kingdom establish the Atrocity Crimes Advisory Group (ACA) for Ukraine.

[12] *See* American University, Journal of Gender, Social Policy and the Law, How Universal is Universal Jurisdiction.

[13] Nicholas Taylor is the Deputy Group Leader for Technology Strategy and Services at the Los Alamos National Laboratory Research Library. He also acts as a temporal web forensics expert and has consulted on many legal matters in the US and other jurisdictions. During this roundtable he was appearing in this latter capacity.

[14] The Wayback Machine is an initiative of the Internet Archive, a 501(c)(3) non-profit building a digital library of Internet sites and other cultural artefacts in digital form. The Wayback Machine has archived 25+ years of web history, is accessible to anyone with internet access, and is considered the foremost institutional web archiving authority. Further, it is frequently relied upon by litigants who seek to present such archives for admission into evidence (particularly in US courtrooms). See Internet Archive, About the Internet Archive.

[15] *See* Internet Archive, Standard Affidavit – which notably limits itself to attesting to the authenticity of the affixed records and not to their probative weight.

[16] *See Federal Rules of Evidence, Rule 201*

Wayback Machine evidence has been more jurisdiction- or judge-dependent. The Internet Archive affidavit and expert witnesses have tended to be the most reliable approaches.

1.4 - Conclusion

These evidentiary standards merit several observations. First, it is clear that the admissibility standards within the international criminal tribunal system and continental Europe (where most investigations are taking place) administer a rather permissive threshold for admitting digital information into evidence at the initial stage, particularly when compared to the stricter approach taken to admissibility in criminal proceedings in common law systems. However, even where evidence is admitted, judges ultimately assess and evaluate the probative value of any particular item of web archive material - i.e. whether it is treated as convincing evidence of the facts it is adduced to prove. But the probative value that is attached to any particular item of web archive material may not vary as considerably across jurisdictions.

Accordingly, even with the ostensible ease with which documents may be admitted, the focus must fall on ensuring that archived evidence will be considered highly reliable so that the Trial Chamber gives it weight in its ultimate assessment. This requires best practices to be implemented for web archives, digital evidence, and all evidence more generally.

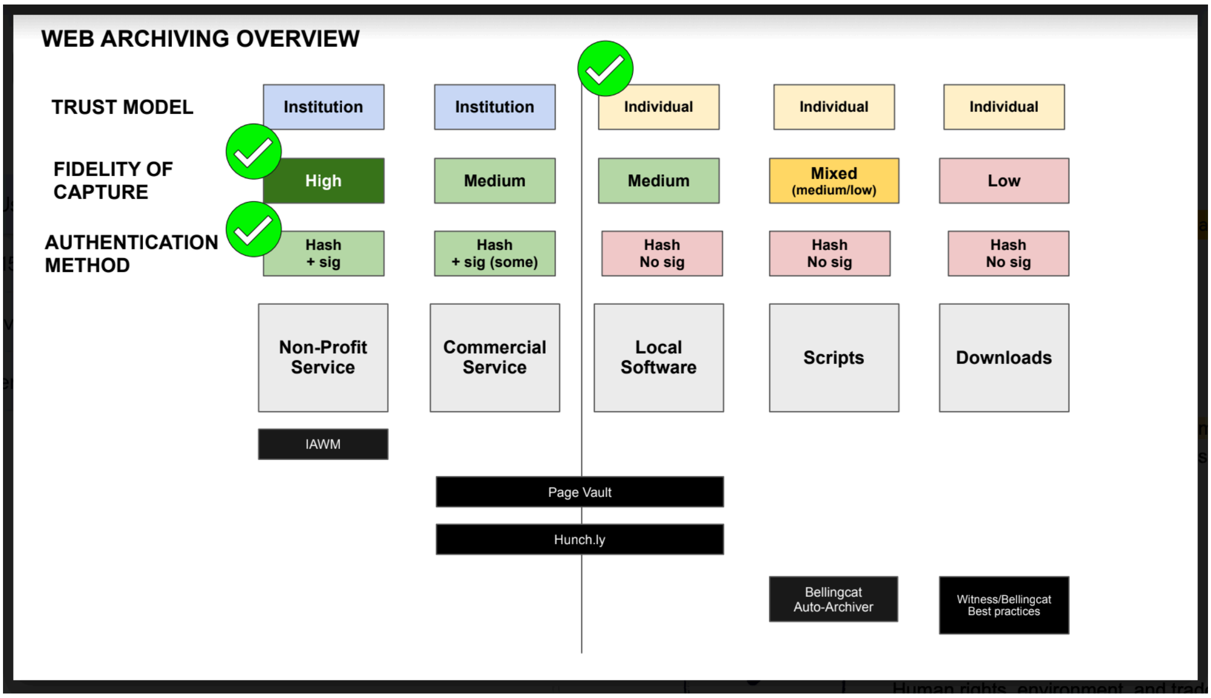
From a practice perspective, it is important that those investigating crimes do not follow these permissive evidentiary practices. Even if some judges may accord low-fidelity web archives probative weight at the evaluation stage of the proceedings now, there is concern that this should change once judges better understand the vulnerabilities inherent in such web archiving practices. One roundtable participant noted that while courts have not approached web archive material with a sophisticated eye to date, this may indeed not be the case for long as both judges and lawyers are becoming more savvy in understanding, and potentially challenging, the admissibility and weight to be accorded to digital evidence.

Finally, while the ICC operates more similarly to a civil law court, this appears to be slowly changing and particularly so relating to evidentiary practices. As time passes, features of both the civil law and common law systems are informally permeating the international criminal tribunal system due to the practice backgrounds of the lawyers and judges. More robust admissibility standards and principles derived from common law courts can be expected to gain increasing traction.

A failure to follow best practices increases the risk that the materials will not be accorded substantial weight. Accordingly, it is necessary to ensure that the highest standards are followed in the treatment of web archives if the materials are to play an important role in international criminal processes.

2- Web Archiving Characteristics

The roundtable discussed a range of web archiving practices, each with its own distinguishing characteristics and challenges. In the sections that follow, **we present a framework for assessing the qualities and limitations of web archives**. This framework – see diagram below – will benefit courts and litigants in understanding how to increase the reliability of a web archive as well as situations where a web archive may be vulnerable to challenge at the admissibility stage of legal proceedings. It will also assist those conducting pre-trial investigations, particularly if they follow the best practices approach identified throughout this section.



WEB ARCHIVING OVERVIEW

TRUST MODEL	Institution	Institution	✓ Individual	Individual	Individual
FIDELITY OF CAPTURE	✓ High	Medium	Medium	Mixed (medium/low)	Low
AUTHENTICATION METHOD	✓ Hash + sig	Hash + sig (some)	Hash No sig	Hash No sig	Hash No sig
	Non-Profit Service	Commercial Service	Local Software	Scripts	Downloads
	IAWM	Page Vault Hunch.ly		Bellingcat Auto-Archiver	Witness/Bellingcat Best practices

It is important to note that none of these characteristics is essential to the admissibility and probative value accorded a web archive. Instead, they represent different levers that each have their own pros and cons as far as each consideration is concerned. Understanding these characteristics can help archivists design effective web archives and assist lawyers in comprehending the advantages and drawbacks of particular web archiving techniques.

2.1 - Trust Model: Institutional v. Individual Web Archiving Practices

One defining characteristic of different practices of web archiving is whether the operator carries out their practice on behalf of an institution, or as a single individual. An institutional link attaches special qualities to the archive itself, as though the institution's reputation and social stature embeds in the work. On the other end of the spectrum, material that is collected by an average person working on their own would probably not benefit from this boost of institutional capital.

The Internet Archive, a prime example of an institutional practice of archiving, has generally established itself as a trusted source by courts for web archives. Accordingly, should the

Internet Archive submit an affidavit in support of a web archive downloaded from their website, courts would likely accept it into evidence.[17]

While institutional archives are generally viewed as credible, one specific concern with the Internet Archive is that it was not created to be used as evidence in legal proceedings[18]. Specifically, there may be limitations on the extent to which an archive retrieved from the Internet Archive can be said to represent precisely what a given webpage displayed at a particular point in time. Such limitations must be acknowledged and with them, possible vulnerability to evidential challenges that may be brought in the future.

On the other hand, individual practices tend to rely on different software services, such as a subscription to Hunch.ly or the recently released Auto Archiver script from Bellingcat, among many others. Generally speaking, when creating a web archive using these tools, archives are downloaded by the individual onto their local computer, or stored by the service used. Trust in these archives fundamentally depends upon the process that the archivist follows and documents, the archiving service that they use, and the identity of the person who runs the archive.

2.2 - Fidelity of Capture

The fidelity of the capture of a web archive is also an important consideration when looking at best practices. From low-fidelity to top-tier forensic quality, we can qualify three tiers of fidelity within existing practices.

An example of a web archiving practice at the bottom of the scale of fidelity (and most likely to fail a proper admissibility challenge) would be taking screenshots and saving them to one's personal computer. While this could technically be constitutive of a web archive, it is of the lowest fidelity (and of likely little probative value in the eyes of a court) because it is particularly vulnerable to forgeries and manipulation given the lack of opportunities to prove otherwise[19] and reliance on one single file.

Moving along the fidelity continuum, *stronger* web archives include not only the visual representation of the information (as in a screenshot) but also its constitutive parts: HTML source code, style sheets, and media assets. Videos, which are technically slightly harder to capture, can additionally be captured by these web archives.

In addition, and still moving towards more-and-more fidelity, the web archive might also include a replay engine permitting browsing of the web archive through a familiar web browser and in situ. Such replay capabilities defend against the ever-changing landscape of web standards and browsers, and permit investigators to place themselves in the technical landscape as it was on the day of the archival – effectively defeating content rot.

[17] For example, in a review of admissibility determinations between 2004-2022 in US federal courts concerning the Internet Archive's Wayback Machine, Nicholas Taylor identified that an affidavit was sufficient in 78% of cases.

[18] However, in Rutherford v. Evans Hotels, LLC (2020, S. D. Cal.) the Internet Archive was found to be more trustworthy than a paid web archiving service built in anticipation of litigation, precisely because the Internet Archive had no vested interest in serving that use case.

[19] To this effect, see below for the benefits of including cryptographic hashes and signatures.

To maximize the likelihood of admissibility of the web archive into evidence, only high-fidelity captures should be archived for later proffering into evidence.

2.3 - Authentication Methods

While the above two sections address how a web archive is made, the method for authenticating a web archive addresses its overall quality and, in particular, how it is preserved. To authenticate a web archive in line with best practices, it is necessary to calculate its unique hash values (integrity), as well as incorporate the use of cryptographic signatures (origin).

A hash is a cryptographic tool that, once used from a piece of data like a video or image, ensures it has not been altered. A hash is created by running data through an algorithm that generates a short alphanumeric representation of the data. Any change to the original data used to create the hash will dramatically alter the hash value. If the present hash matches the past hash of that data, it is mathematically proven that the data has not been altered or manipulated, and that the data is authentic.

In addition to hashing, cryptographic signatures permit the setting and demonstration of the origin of a claim, i.e. who archived this web page. They work in two directions. First, they allow someone to sign a claim with a private, securely-held key. Second, they allow anyone to verify that the signature was done by the claimed signer without requiring the aforementioned private key.

The use case presented during the workshop used domain names and HTTPS, which are pieces of cryptography functioning in the way outlined above: a user is given guarantees they are browsing their bank's website without needing the bank's secret key. In the case of web archiving, an institution would grant explicit consent to a researcher using a mechanism to sign the material they produce with the domain certificate. After the fact, it is perfectly possible to prove it was indeed someone at this institution who captured the archive, and that the institution must have consented to it in the first place.

Courts already find the use of cryptography reliable in other areas of their evidentiary practice. For example, electronic signatures on contracts and other legal documents have been found to be valid and authentic for more than twenty years. Raising awareness of the security that signatures offer and their utility should be prioritized.

3- Web Archiving Best Practices

After reviewing a web archive's general core characteristics, the roundtable discussed a set of best practices (or a 'wish list'), which sets forth a picture of the ideal web archive for use in court, drawing on the requirements of the Berkeley Protocol on Digital Open Source Investigations.[20]

3.1 - Ideal Web Archive Characteristics

The ideal web archive:

- Could have been produced by anyone, meaning that the material was publicly accessible (i.e. not privileged or obtained through deception) and the tools to produce it are available to everyone (i.e. not proprietary). This could be either or institutional actor or an individual (see characteristic No. 1 above);[21] and
- Is of high fidelity, meaning it was carried out by a tool that preserved much if not all of the original material (see Characteristic No. 2 above); and
- Includes the content itself, its surrounding metadata, the metadata of the web scraping software, its hash value, and the signature of these hashes authenticating it to the author (see characteristic No. 3 above).[22]
- The aforementioned hashes and signatures must be preserved, that is to say stored securely and made available for the long term, as would the content itself.

With such practices in place, a secure web archive (or extracts from it) can be presented in a way that maximizes the prospects of a court admitting it into evidence and affirming its probative value.[23]

3.2 - Establishing Clear Methodologies between Lawyers and Tech Professionals

Lawyers and tech professionals should work jointly to ensure observance of and insistence on best practices so that digital archive material is of the greatest possible use in international criminal trials. The best practices outlined above can be supplemented by reference to independent standardization guidelines such as the International Organization for Standardization's document entitled "Information and Documentation - Statistics and Quality Issues for Web Archiving".[24]

[20] *See* Berkeley Protocol, Berkeley Protocol on Digital Open Source Investigations: A Practical Guide on the Effective Use of Digital Open Source Information in Investigating Violations of International Criminal, Human Rights and Humanitarian Law ('Berkeley Protocol'), p. 53 (retrieved 31 Oct 2022).

[21] As long as best practices are observed, individually-downloaded archives can be trustworthy and easy to use. Indeed, if individual web archiving practices were widespread and well-understood, they could even be used by ICC investigators themselves in conducting pre-trial investigations. It would also lead to a marked increase in 'citizen' investigators who can collect and submit such archives to a range of legal mechanisms.

[22] *See* Berkeley Protocol, p. 59 for a description of all elements that can be captured (including those listed in these bullet points), thereby providing a strong foundation for establishing the authenticity of a piece of digital information.

[23] Finally, while not discussed above, it is important that web archives also follow best practices in the preservation of information. Such practice should involve conformity with the Berkeley Protocol, pp. 60-62. Starling Labs is active in promoting such preservation strategies to ensure data integrity and concomitant proof for the consideration of an adjudicative facility when deciding whether to admit a document into evidence and accord it probative weight at the time a decision is made.

[24] See International Organization for Standardization, Information and Documentation - Statistics and Quality Issues for Web Archiving, ISO/TR 14873:2013(en).

To maximize the likelihood that a web archive will be admitted into evidence, archivists should establish clear, detailed methodologies to facilitate admissibility. In proving that a document is authentic (i.e. it is what the moving party says it is), the law needs to know where something comes from, how it was procured, who procured it, when it was procured, the process that the individual followed, the chain of custody, and the ability (or inability) for the archivist to adjust or otherwise modify the webpage that they are archiving.

A clear methodology should also include the identity, qualification, and training of the person conducting the research, as well as the relevant web collection protocols that were observed (e.g. Starling Framework, ISO, etc). A description of the relevant hardware and software being used to archive the web page is also important (e.g. Internet Archive, Hunch.ly, Page Vault, etc). Archivists should also identify their process for selecting certain websites (and articles within websites) and assess them for credibility and trustworthiness, both in terms of the content and its susceptibility to manipulation. Efforts to identify the author and steps to corroborate their identity should also be indicated.

Storage methodology should also be clearly detailed. This includes describing the protocols in place to ensure against corruption, hacking, and other situations that could arise which could damage the archives. This and other relevant demonstrations will be used to establish that no information has changed during the storage process. This should be recorded in a chain of custody that documents who has handled the document.

4- Trial Techniques and Practices to Improve the Likelihood of Web Archives being Accepted into Evidence

4.1 - Relying upon Different Practices for Seeking Admission of Web Archives into Evidence

There are a range of steps that lawyers can take to improve the chances that archived data will be accepted into evidence. At the ICC, this could include establishing a clear pre-trial protocol at the beginning of a criminal trial which will clarify the standard and practice level expected when seeking the admission of a piece of digital information into evidence. Such clarity will serve as a 'pre-screening' standard that - so long as it is diligently followed - should facilitate evidentiary admissibility. Such pre-trial protocols are common in the international criminal tribunal system.[25]

Parties to the proceedings should also consider the propriety of pushing for a stipulation between the parties as to their view concerning the admissibility of web archives. In principle, both parties are likely to seek to enter certain information into evidence. If they agree to the authenticity of each document (but not the evidentiary value of the underlying content contained in the web page) in advance, it will markedly improve the admissibility determination when the document is actually being sought for entry.

[25] *See* example here.

It is not advisable to rely upon judicial notice as a vehicle to ease the admissibility requirements for web archives. While nearly everything presented in the court case needs to be proven in that particular case if it will be used against an Accused, judicial notice is an acknowledgement by the court that some matters do not need to be proven because their authenticity cannot be reasonably questioned. Some courts have taken judicial notice of web archives, and a subset of those have been overturned on appeal.

The reliability of web archives, particularly the Wayback Machine, raises interpretative concerns that courts should address carefully when considering judicial notice. The primary issue is not the authenticity of the archived assets, which are reliably preserved and attested to by the Internet Archive's affidavit process, but rather the potential creation of temporally inconsistent composites. These composites, formed by combining assets archived at different times, can present challenges for proper interpretation and may affect the probative weight of the evidence. While the affidavit confirms the authenticity of the underlying records, it does not involve a specific analysis of the content or its temporal coherence. Consequently, the replayed composite snapshots cannot always be taken at face value as accurate representations of what was displayed at a specific point in time. Legal professionals should therefore contextualize the archived material rigorously to mitigate the risk of misinterpretation. Nonetheless, the Internet Archive and its processes for collecting, storing, and presenting content remain trustworthy and invaluable tools for litigators, especially in capturing relevant content that might otherwise be lost. Courts can balance confidence in these archives with the need for careful scrutiny to ensure accurate and fair use of this evidence in legal proceedings.

Finally, if a particular web page is of the highest materiality to proving the allegations of a particular case, it would be useful to obtain an affidavit describing the process and/or call a witness to authenticate the archive.

4.2 - Consider Changes to Technical Terminology

Finally, perhaps the most important practice when seeking to have web archives entered into evidence - and one mentioned several times during the roundtable - would be changing the complex terminology of web archiving best practices into language that is more suitable for the legal world. Mr. Dotan presented a variant of this during the roundtable when he analogised the collection of forensic evidence at a crime scene using sealable evidence bags containing detailed information about the item and its chain of custody on the outside of the bag to that contained in a web archive. Such analogies can assist courts in improving their understanding of the technical nature of web archiving and give them assurance before admitting documents into evidence and according them the highest probative value possible.

On the technical side, tech experts will need to be able to explain the role of and importance of hashing and cryptographic signatures in authenticating digital records, and illustrate their ubiquitous daily use in other fields in which trust and authenticity are essential (e.g. in connection with financial transactions and secure internet browsing). To allay any doubt or ambiguity, when tech professionals explain their method it may be necessary to analogise their approach to authenticating web archives in a manner that lawyers and judges can understand, such as the image of the physical plastic sealable bag above, and how its features map to those of hashing and signing a digital item.

The same applies to terminology that may contain phrasing that is not typically helpful in a courtroom. For example, the term “non-trusted source” is understood positively in the blockchain industry. Such a notion being understood positively is anathema to any judicial system, as the notion of trust is how a judge bases all decisions. If they trust the web archive, the documentary evidence, or the witness, they will deem the evidence reliable and accord it weight when making decisions on the culpability of an Accused.

As an additional example, the value of decentralization should not be taken for granted. Instead, a party to the legal proceedings may be better off focusing on the benefits of such decentralized systems – spreading the risks inherent in web archiving to many different storage systems, ensuring that there are multiple points of failure in the event of a technical malfunction, and the use of cryptography (the gold standard for secure transactions on the internet for decades).

4.3 - Furthering Trust in Web Archives: Institutional Corroboration and User Protection

Finally, to improve on the process of entering web archives into evidence, Starling Labs have created the concept of “Witness Servers” as an additional layer of self-corroboration for web archives. A Witness Server is a service, hosted and run by an institution, which carries out web crawls on-demand on behalf of individuals conducting web archiving activities. Several Witness Servers act in concert on the instruction of a web archivist and simultaneously capture the same web page. Such an approach addresses the possibility of a webpage having slight variations depending on locale (and many other potential anomalies) and works to otherwise corroborate the contents of a website through a replication process that validates the contents of a web archive from several different locations.

Starling Labs has recently released a Concept Note and Call for Contributions to gain further information from relevant legal experts in the field. Upon the conclusion of such a review, changes will be made and a full release will be publicized.

Conclusion

The evolution of web archiving has become an essential tool in the pursuit of justice, particularly in cases requiring robust digital evidence. As this white paper highlights, effective web archiving practices must balance technical rigor with legal admissibility to ensure their probative value in courtrooms worldwide. By incorporating best practices—such as maintaining high-fidelity captures, utilizing cryptographic authentication, and fostering collaboration between legal and technical professionals—archivists can significantly enhance the reliability and utility of web archives as evidence.

Moreover, the workshops presented in this paper underscore the importance of bridging the gap between technology and law. Establishing clear methodologies, adapting technical terminology for legal contexts, and integrating innovative tools can further bolster the trust and accuracy of web archives. As courts, investigators, and litigators increasingly rely on digital evidence, it is imperative to refine archiving practices to uphold the principles of justice, ensuring that even ephemeral digital content can serve as a cornerstone for accountability and truth.