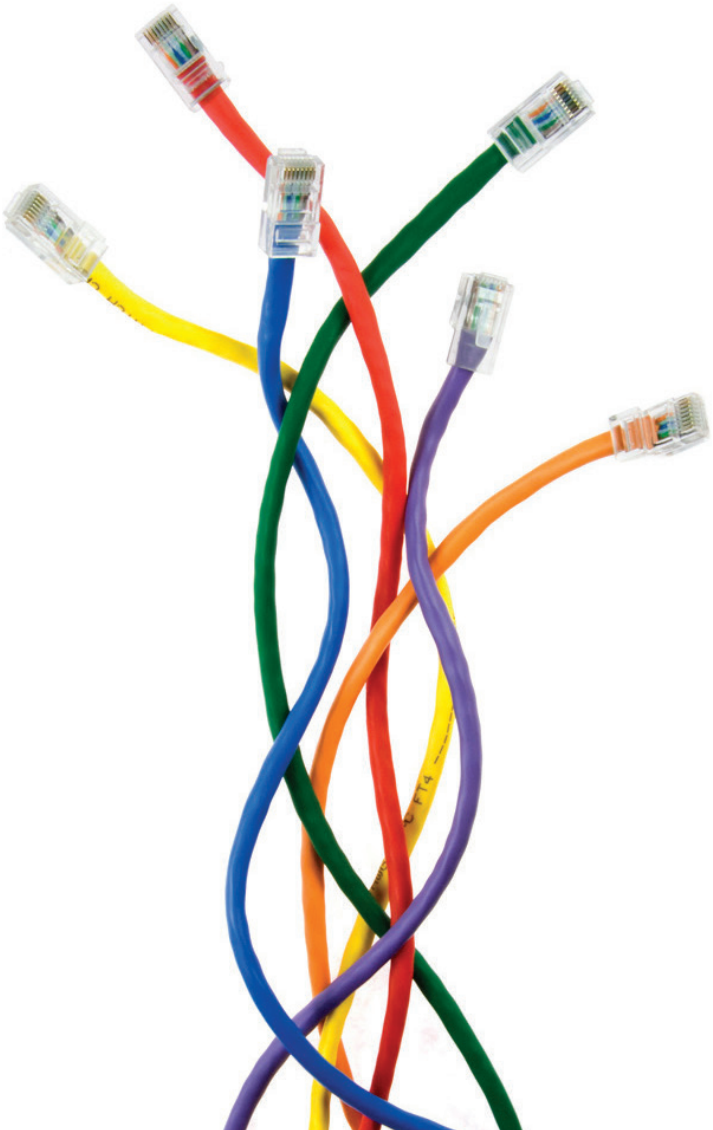
Deloitte.

# WARC this way

## eDiscovery and the ISO 28500 standard for web content collection and preservation



**With the ever growing prevalence of transactions, communications and operations conducted solely on the web, it is important for the eDiscovery community to examine standards for web content collection and preservation. This article discusses the ISO 28500 Web ARChive (WARC) standard for web content collection and preservation and its evidentiary implications for electronic discovery.**

Let's begin by saying that reasonableness is the real standard in most things, including eDiscovery. Put another way, if there is a standard for a given task, are its risks, costs, timing and workflow such that the standard should be followed or not? Courts using this standard will look at both the ultimate decision, and the process by which a party went about making that decision. Of course, each case, each court and each client is different, and adherence to eDiscovery standards and leading practices varies with each risk profile, but it is safe to say it is *usually* a good idea to adhere to standards and leading practices. The ISO 28500 WARC file format is the internationally recognized standard for website preservation and therefore, a good starting place for ediscovery practitioners.

Web content is anything delivered over Hypertext Transfer Protocol (HTTP), including web sites, social media sites, intranet sites, Wikis, blogs, and other similar sources. Web content preservation and collection is not new, but the inclusion of web-based content in litigation and regulatory investigations is becoming more the norm rather than the exception. For example, in the case, *E.E.O.C. v. Original Honeybaked Ham Co. of Georgia, Inc., No. 11-cv-02560-MSK-MEH (D. Colo. Nov. 7, 2012)* involving allegations of sexual harassment, a hostile environment and retaliation, the court granted, in part, the Defendant's Motion to Compel and ordered discovery of the class members' social media (Facebook), text messages and email. Prior to the ruling, the court indicated that class members had utilized "electronic media to communicate" about potentially relevant topics and described that content "as though each class member had a file folder entitled 'Everything About Me,' which they have voluntarily shared with others" and that if there was relevant information that could lead to the discovery of admissible evidence within this folder, "the presumption is that it should be produced." The court further reasoned that the fact that the evidence resided on the web, "[i]s a logistical and, perhaps, financial problem, but not a circumstance that removes the information from accessibility by a party opponent in litigation."

If you happen to work in web archiving for the Library of Congress, or a national library of virtually any country in the world, you are likely aware that the ISO 28500 WARC format is the standard for web content collection and preservation.

### About the WARC standard

The WARC standard was developed by the International Internet Preservation Consortium (www.netpreserve.org), and it serves as a benchmark for the proper methods and procedures for collecting, preserving and storing web content.

- It dictates that web content be preserved in the exact, native format it was presented when it was live on the web. In other words, it must be designed primarily to store the native content (payload) and all the protocol data used to collect it (metadata). Transformations of the web content are permissible, but only as supplements to the native.
- When rendering captured web content, it must be identical to the live site at the time of capture: links should work, animations and videos play, forms function, calculators calculate, etc.
- It mandates an audit trail to the original content, essentially showing a documentable path that can be traced back to the original web site.

If this sounds familiar to other accepted evidentiary and eDiscovery practices, it should. It is very similar to leading practices for other native files in eDiscovery. If you collect an Excel spreadsheet, you would not transform the file format, and you would expect the Excel functions to work the same when you later open the file. And, you would expect the collection to be done in a forensically sound manner.

Fed. R. Evid. 901(a) states: "The requirement of authentication or identification as a condition precedent to admissibility is satisfied by evidence sufficient to support a finding that the matter in question is what its proponent claims." To be admissible, web content evidence, like any other form of evidence, must also be relevant, authentic and admissible. Authenticity may well be the most common challenge to web content evidence because of the ease with which the content may be altered, either intentionally or due to sub-standard preservation and collection methodologies.

As mentioned above, the WARC standard specifies that web content be captured in native format i.e.; a file saved in the format as designated by the original application used to create it (e.g. a DOC or DOCX file created by Microsoft Word). Again, native format is familiar to the eDiscovery community, but with web content there are a few complications. Transformation of web content info .mht, .htm, .pdf or other file formats as part of the collection process should be a complement to the native format content, not a replacement. It is acceptable to create a PDF or .mht rendering once the native content has been captured and preserved properly but only if you maintain the original native web content.

It is important not to confuse content viewed through a browser such as Internet Explorer, Chrome, Firefox, etc., with truly native web content – just like you would not confuse a Word document in a native file viewer with an actual Word document.

### Another web wrinkle: improper archiving and the chance for spoliation

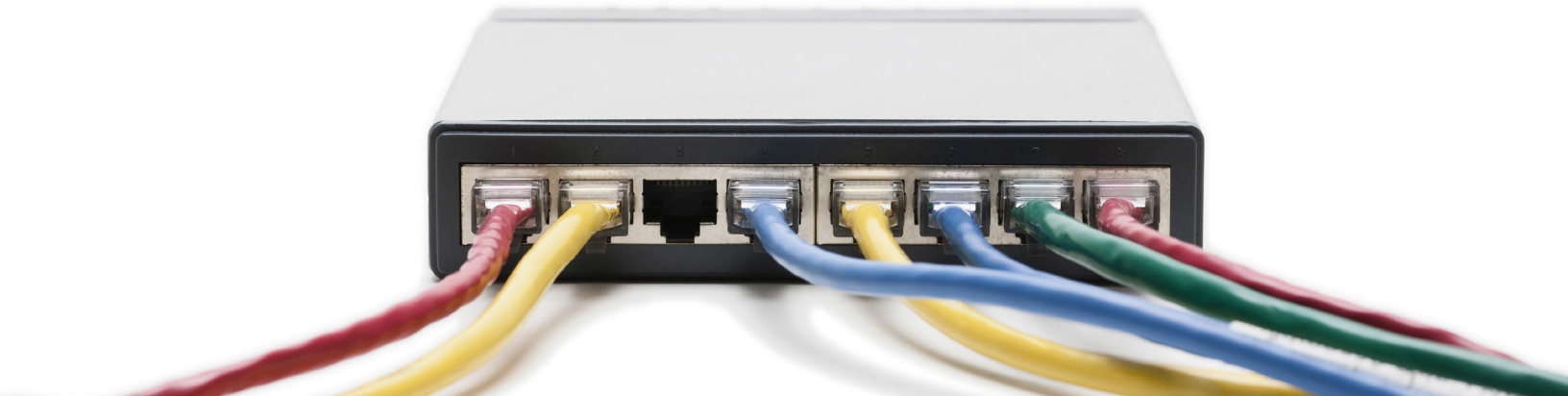
Another important requirement of the WARC standard regards the actual "archive" itself. The archive must be buffered from any contact outside of the archive, i.e.; the live web. Modern web content is highly interactive. What that often means is that a web site is configured to call a server, or another live web site, in order to fully display content to the end user. This allows a website to constantly refresh content by calling up news stories, images, videos, etc., from sources other than the server where the website resides. This is where the danger lives from an eDiscovery perspective. If you are viewing an archived web site that has presumably been "preserved" and some of the content you actually see on that site has been updated from the live web, has it really been preserved? If any of the data that has been updated or replaced could be considered relevant to the matter, then it could be claimed to be spoliation.

For example, consider the video on your client's site that is actually served from YouTube. If you preserve the site, but the video is still accessed via a link to the live YouTube site, and YouTube later changes the video or takes it down, you have a situation that is somewhat similar to collecting a macro-enabled spreadsheet without collecting the macros. How often do postings and photos on web sites such as Facebook, Twitter and similar social media sites change? Is not the same true for stock prices and other financial information displayed on line? Oftentimes, this content originates from another source and every time the site is visited it "calls out" for updated content.

A proper archive that has been built around the WARC standard prevents links to live web content and preserves all content as it was at the time of collection. All content necessary to meet the requirement of identical display must be collected and preserved in the archive. Boundaries on the collection must be established prior to collection.

### The WARC standard and APIs

What about web content collected via an API (application programming interface)? An API is a method that many companies make available to make it easier for other applications to exchange information with the company's application. For instance, Facebook offers an API so that other applications can import data and information from its site.



By definition, an API takes data out of its original context and format. Whether that is good or bad from an eDiscovery standpoint depends on many factors including the nature of the case, the content itself, and the client's risk tolerance. But, collection of web content solely via APIs is not a comprehensive native format collection. The ISO 28500 WARC standard would consider data acquired via APIs as supplementary data, or transformations, of the original web content. APIs transform web content from its original state and on its own, this may well open up the evidence to authentication challenges.

In his memorandum opinion in Lorraine v. Markel Am. Ins. Co., 241 F.R.D. 534 (D. Md. 2007), Magistrate Judge Paul Grimm from United States District Court for the District of Maryland, wrote that "considering the significant costs associated with discovery of electronically stored information (ESI), it makes little sense to go to all the bother and expense to get electronic information only to have it excluded from evidence or rejected from consideration during summary judgment because the proponent cannot lay a sufficient foundation to get it admitted." While using an API to collect web content may seem convenient and less costly, the evidence may be more at risk of being inadmissible or it may require costly testimony to ensure authentication.

Another reason web content collection exclusively via API is risky is that after web content has passed through an API, it may not be rendered in an identical way to the original web content and user experience. As with other eDiscovery data sources, unavailable native content, or content that has been substantially altered from native form, often means a more difficult and more costly review andproduction process.

But, that does not mean that APIs are bad, not at all. Compare them to eDiscovery practices around emails. For a relatively simple dispute, lawyers may be completely comfortable with asking a client to simply forward relevant emails because the risk is low. Web content is already treated the same way: in many cases, it is acceptable to save a web page as a PDF, or to use an API-based solution to collect posts and comments from Facebook, for example. However, as with email evidence, some cases warrant a more rigorous process. In cases where the standard of evidence is higher or the risk is greater, it is useful for eDiscovery practitioners to be aware of the WARC standard to avoid questions about the authenticity or completeness of evidence.

Web technology changes very quickly, and the cycle from new to adopted to obsolete seems much shorter these days. In 1999, a district court cautioned against relying on data from the web as "voodoo information." In St. Clair v. Johnny's Oyster & Shrimp, Inc., 76 F. Supp. 2d 773, 775 (S.D. Tex. 1999), the court wrote, "While some look to the Internet as an innovative vehicle for communication, the Court continues to warily and wearily view it largely as one large catalyst for rumor, innuendo, and misinformation… [f] or these reasons, any evidence procured off the Internet is adequate for almost nothing."

Times have changed. In 2014, as web evidence becomes even more common in litigation, there is little doubt that the preservation of web content as evidence is sure to be a well-examined topic into the future.

**Contatct**

**Richard Vestuto**
Director | Deloitte Advisory
Deloitte Discovery
Deloitte Transactions and Business Analytics LLP
rvestuto@deloitte.com
+1 212 436 2044