hanzo

WEB MEDIA ARCHIVES, AND HOW TO DERIVE VALUE FROM THEM

WHAT IS AN ARCHIVE?

In its simplest description, an archive is ‘a collection of historical documents or records providing information about a place, institution, or group of people.’ Archives frequently take the form of collections of information/records i.e. letters, reports, registers, census data and in more recent times digital data. These records are often primary sources and differ from a library, typically composed of predominately secondary information sources.

WHY SHOULD THE WEB BE ANY DIFFERENT?

Web archiving is the process of collecting anything viewable from within a web browser and ensuring that this information is preserved in an archive for future researchers, historians, corporations and the public.

The Internet is great at showing the now, the current product, the current price, logo, video or other IP. The Web's ephemeral content is continually updated, restyled or replaced. It is a vast unstructured database of evolving content. URL's, domain names, company logos, addresses, information, portfolios and so on may exist for only a year or two before they're updated meaning the old information has been lost! Now, the fluidity of content is increasing as the shift from the traditional email and website model to more dynamic systems such as social media, video, 'Wikis' and virtual offices—meaning the rate of loss of content and information also increases.

Backups of web-servers are maintained for disaster recovery purposes but little more and again they only mirror what's live. Older information is continually discarded to make room for the new. Shouldn't a company's web output be used to provide the same insight and information about the business that created it? Shouldn't it be protected in the same way?

Because server-side technologies are continually updated, keeping legacy content online is increasingly difficult as it becomes incompatible with the latest versions of say PHP, JavaScript or MySQL, or risk being compromised when older vulnerable technologies are placed online at the mercy of hackers. This is perhaps the main reason why the Web is typically overlooked for archiving as it's not widely known that client-side capture that respects digital preservation policies will negate these issues and ensures content can be preserved securely and indefinitely using open standards that won't end up being inaccessible.[1]

Digital preservation ensures that digital information that has continuing value remains accessible regardless of the challenges of media failure and technological change. The goal is the accurate rendering of authenticated content over time. According to the Harrod's Librarian Glossary, digital preservation is the method of keeping digital material alive so that it remains usable as technological advances render original hardware and software specification obsolete.[2]

Companies use analytical software or SEO techniques to ensure their current site is receiving the attention desired but how do they draw usable data from what it was and compare it to the now? Hosting incremental copies of web content is challenging and expensive as web servers are updated or replaced rendering the code they were written in obsolete. A marketing department may keep a record of campaigns and promotions but are unable to see how they appeared on the web site when it's no longer there. They need to retrieve statistical or analytical data to determine the success of the content and how it was displayed.

But if we are to treat our web archives as this 'collection of historical documents or records that provide information' as per the description earlier then the tools to extract and analyze the data they contain have to be up to the task. Archives should be searchable, referenced. Live content comparable with that of the archive. How much of your information is no longer available on-line? What information is truly in there? If the live content to be captured is dynamic then the capture window is short, if it exists now then it should be captured now and time-stamped to evidence when it occurred.

The thousands of records known as WARC files (Web ARChive) generated from the archiving or 'crawling' process accurately record the content at the time of capture, for them to be legally defensible then they should be accurately indexed and stored securely to prevent any possible tampering.

Capturing content from elsewhere on the Internet may also prove to be a useful data source; social media, competitor's sites, rogue tradesmen, fraud, IP infringement or other items of interest could be recorded and treated to the same analytical scrutiny.

WEB CURATION

Web curation, like any digital curation, entails the ability to demonstrate:

- Certification of the trustworthiness and integrity of the collection content
- Collecting verifiable web assets
- Providing web asset search and retrieval
- Semantic and ontological continuity and comparability of the collection content

Once it's decided what to capture and why collect the material along with supporting metadata in a repeatable and testable way to ensure authenticity and avoid misinterpretation or second-guessing.

Often these key functions are either missed or fail to be sufficiently robust to withstand scrutiny, if it's worth recording then it should be done with these considerations or it's potentially worthless![3]

WHAT CAN YOU DO WITH A HANZO ARCHIVE – OUR SOLUTIONS?

Businesses often wish to keep their web content and media, frequently deciding to archive their own content for various legitimate purposes:

- Meeting specific regulation e.g. SEC rule 17a-4 on storing electronic broker-dealer records.
- Corporate heritage and knowledge e.g. use in corporate museums or a historical record of online publication;
- General governance, regulatory or compliance reasons i.e. the golden source of truth on what was on their web systems at that particular time;
- Legal purposes i.e. defending or making a claim such as eDiscovery or litigation response;
- Insurance i.e. meeting conditions of a policy or keeping work product for future use;
- Others.

USER CASES FOR LEGALLY DEFENSIBLE DIGITAL ARCHIVES

Some of the use cases we have come across are listed below but the potential list of applications is virtually limitless. That said the start point is the same, a business always needs to be capturing and its storing web, social media and collaborative content defensibly.

Intellectual Property Protection (Patent, Trademark, Copyright, Know-How, Business "Get-Up" and Passing Off)

A legally defensible archive can be used to demonstrate when an idea or product first came into existence and indeed allow a corporation to prove it was the first to invent; a vital part of the patent process in the USA. An irrefutable record of when and who was first to create this can be hugely valuable to a company.

Frequently companies suspect their IP is being used without their permission (or suitable license) by a competitor. Proving such an assertion can be a challenge especially when infringing evidence is at risk for the quick removal from public view. Obtaining a legally defensible copy of the incriminating evidence that links it to the competitor is invaluable in preventing an IP breach. Examples include the reknown Apple v Samsung patent dispute.

Discovery of the activities of another can be used to stop content that belongs to one corporation from being used elsewhere without the creators/owner's express permission. This has implications for brand protection and preventing employees from taking corporate secrets or knowledge with them when they change employers.

This platform can be used to assist a company in proving that the content belongs to them and making an application to a regulator, enforcement agency or lawyer to have the content removed or acted upon. This will help in stopping the violation and/or prevent further leakage of critical business know-how.

Third Party Due Diligence (internal and external sites)

Due diligence must be risk-based and should always be proportionate to the risks posed by the specific third party e.g. its location, type of transaction/business contemplated or associated persons.

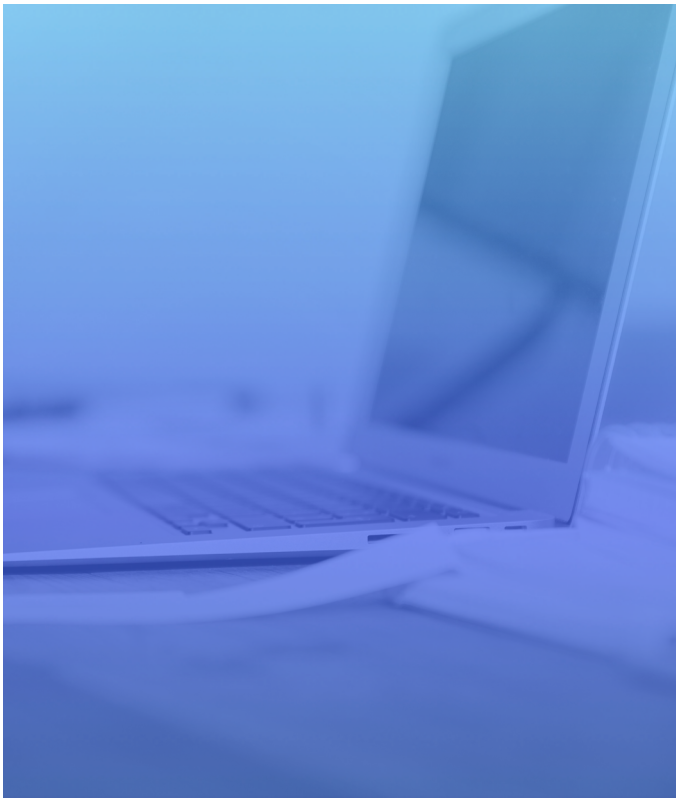
Companies which conduct business internationally face increasingly complex legal and reputational risks. There is mounting pressure from regulators, enforcement agencies and civil society, as well as a dramatic increase in levels of business being carried out in higher risk jurisdictions. Many risks go unseen as they lie in a relationship with a third party, and traditionally due diligence has found it difficult to understand these risks.

Today it is possible for a business to perform third party due diligence checks and to maintain this data as a legally defensible record of what information was available when they entered into the business relationship. Clearly identification of a "red flag" gives the corporation time to pull out of the arrangement, whilst equally a "clean bill of health" can be stored and referred to at a future date as evidence of proper checks and procedures being in place.

It is not uncommon for companies to update these third-party evidence archives regularly in response to new information becoming available. Unindexed sites that do not appear on a web search are also accessible; these frequently hold useful information and can be captured to ensure they are preserved.

WEB INTELLIGENCE

Make use of the world's largest unstructured data source; the web (and query the fastest growing social media platforms to augment this information). The Web is full of data, pulling out the relevant data and making sense of it provides a huge advantage to companies. The ability to query, collect, store and iterate/refine a process provides a mechanism for companies to gain an insight into their markets, competitors and keep a record of their observations before the data is removed, changed or amended. Keeping an exact copy of the Twitter feed for the president of the USA or cross referencing a prospective employee's LinkedIn profile through their claimed publications and/or employees can apport valuable information. Organizations can use this as a tool to defend their brand and ensure their reputation remains unsullied.



DATA PROTECTION AND COMPLIANCE

Companies often worry about their customer's personal data being accessible in error or being at risk in the event of a breach. The ability to "sweep" the external systems and report on the location of personal data is valuable from a compliance perspective. Additionally, most companies accept that they will be hacked and want to understand what data is likely to be visible ahead of this happening so that they can remediate before the breach occurs. Maintaining an archive allows this process to become a normal business process and allows a demonstration of the safeguards that were followed if company policies are queried by a regulator.

Equally important is the ability to prove that upon detection of data in the wrong location remedial action was taken and the issue resolved. Furthermore, the ability to demonstrate categorically what terms and conditions were in place at a certain time and accessible to staff can prove invaluable in the legal cycle. Higher Education providers need to demonstrate course content and regulations at a point in time to comply with their obligations under the Companies and Market Authority.

SECURITY AND LAW ENFORCEMENT

How do you know what is on your web presence? Web pages are not just external. There is normally an internal intranet, which contains a large amount of sensitive data. If a page has been compromised, could sensitive information have found its way onto one of your sites without being noticed?

Information security and the safety of a company's systems is a hot topic at present with a number of high profile breaches in the news. It is not uncommon for such breaches to go unnoticed for long periods. Understanding what data resides across your systems can provide invaluable insight into how safe your systems are. Being able to identify and remediate pages that have been compromised is a vital tool in a company's armory. We often find malicious code, rootkits and other malware on the sites we scan. Other common sights include old code that has not been removed but merely obscured by a new format; this data is still accessible and may be damaging. You need to know it is there in order to remove it.

Tracking what users are posting in chatrooms and monitoring what users are saying about your business in public can provide valuable insight into who is talking to whom and why. Combine this with an IP protection strategy and with other more covert intelligence methods to build up a picture of the activity. This can be complicated by the use of third-party platforms in an attempt to conceal who is doing it but these are often web-based and can be brought into scope for collection and legally defensible preservation prior to analysis.

Additional use cases that users can exploit from their archive include using it as a:

- Document store to capture all public-facing content in a searchable and examinable way.
- Open format digital library to ensure it can always be accessed or access can be provided to the public as part of legislation. This minimizes the need to perform complex and expensive data migrations and conversion to proprietary file types.
- Source for translation which can be built in at the source but could also be automated across an archive to archive to ensure it can be understood by all users.
- Monitoring tool by linking archive and statistical data for site visits, you can see how the content has changed and what influence it had on visitors, for example.

By exploring the value of an archive, the information it could contain, and how to make use of it prevents it from simply being a vault of forgotten information and makes it the go-to record of choice to prove/demonstrate what was done and on what basis. As the WARC archive format is a future-proof open standard that does not lock you out of your data —you can extract the value when and how you need to.

It is for the problems and considerations highlighted above that Hanzo's platform was conceived and designed, and why it is the market leading method of capture. Combine this legally-defensible archive and add the analytical tools Hanzo provides and what you have is a powerful mechanism to protect and grow your business.

Sources

1. https://en.wikipedia.org/wiki/Open_Archival_Information_System
2. https://en.wikipedia.org/wiki/Digital_preservation
3. https://en.wikipedia.org/wiki/Digital_curation

To learn more, click below to request a personalized demo.

REQUEST A DEMO