MirrorWeb

A GUIDE TO

Website Archiving

mirrorweb.com

Introduction

Across all sectors, organisations are steadily publishing more and more content online.

Websites are used to sell products and services to clients, to publish and share sales and marketing collateral, and - in the case of government - as a primary channel of communication with members of the general public.

This presents a new set of challenges. Financial firms, for instance, must fulfil a range of compliance obligations in order to use website content and other forms of electronic communication to sell to clients. Elsewhere, some organisations may identify a need to keep legally admissible records of their website content for dispute resolution purposes - and others might wish to preserve website pages simply for their cultural or historical significance, or as a matter of public record.

Website archiving is the only way of preserving website content in a form that enables it to be used as it was at particular point in time.

It is the only means of creating and maintaining a stable, time-stamped, verifiably authentic and independent version of web content. As the archives are independent, they are completely separate from the original website architecture and will only include the elements that were live at that time, portraying the archive in as close to original form (as it was at that point in time) as possible.

There are many reasons an organisation should archive its website but in all cases they must ensure their archives are complete, secure and legally admissible.

In this guide, we look at some of the specific challenges around website archiving in three different sectors - financial services, the public sector and brands - and how the MirrorWeb solution can help.

Why Website Archiving is Important

Website archiving can help companies in the financial, public and retail sectors store immutable records of their web pages to help ensure the following where relevant:

Compliance - Regulated companies and firms must record and retain all electronic communications under MiFID II, FCA, SEC and FINRA rules. MiFID II, for instance, states that the recorded electronic communications must be:

- **Complete** - An organisation must be aware of all types of electronic communications that are used and by whom. In addition to this, they must have a system and processes in place designed to capture and retain all records of those communications.

- **Accurate** - An organisation must be fully confident in the recorded electronic communications' content and metadata which can demonstrate the exact dates and times that anything took place.

- **Quality** - An organisation should be able to reproduce records of electronic communications in as close to their "original form" as possible.[1]

Legal admissibility - Companies could be required to provide authenticated evidence of electronic data when in court. The records must demonstrate that the data has been stored in a format that is unalterable, when it was archived and that the record has not changed since being archived, as described in the following regulations and rules:

- **Federal Rules of Evidence** (rule 901)[2] - the requirement of authenticating or identifying an item of evidence.

- **The Code of Practice on Evidential Weight and Legal Admissibility of Electronic Information** (BS 10008:2014)[3] - ensuring the authenticity and integrity of electronic information.

- **SEC rule 17a-4**, which requires firms to archive electronic business communications in non-rewritable and non-erasable (WORM) format.[4]

Archiving website data for dispute resolution and eDiscovery purposes can help ensure that the records are non-refutable and are a true reproduction of the content at that time.

Protection of IP and brand assets - Brands have a clear incentive to keep a long-term record of their activity to inform future campaigns. However, as more business activity occurs online, they will be continuously creating and publishing large amounts of digital content at speed which can be difficult to keep track of. Website archiving can be carried out on a regular basis, and with unlimited cloud storage and the ability to archive large data sets, it can ensure that nothing of value is lost.

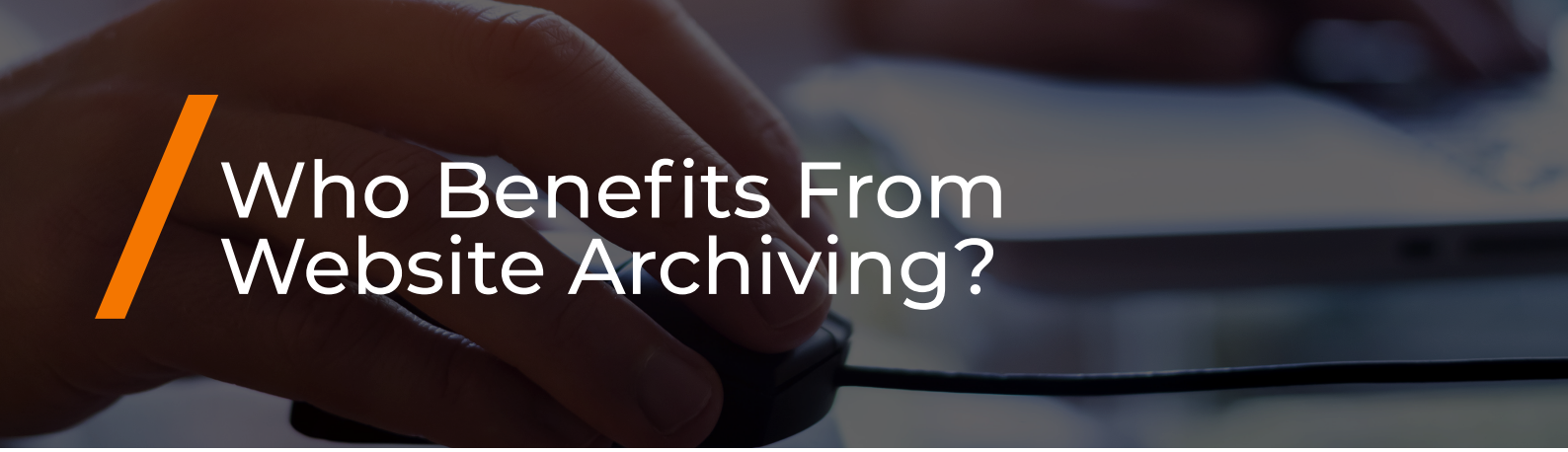
Preservation of records of cultural and historical significance - Public-sector organisations and archivists may require the preservation of culturally important website content for instant access for historical public record data. Website archiving is the ideal solution to preserving large amounts of data and storing in an unalterable WORM file format.

[1] "A Guide to Electronic Communications in MiFID II, Article 16 - MirrorWeb." 22 Dec. 2017, https://www.mirrorweb.com/blog/mifid-ii-electronic-communications-recording-article-16. Accessed 7 Mar. 2018.

[2] "Rule 901. Authenticating or Identifying Evidence - Legal Information ...." https://www.law.cornell.edu/rules/fre/rule_901. Accessed 7 Mar. 2018.

[3] "BS 10008 Electronic information management | BSI Group." https://www.bsigroup.com/en-GB/bs-10008-electronic-information-management/. Accessed 7 Mar. 2018.

[4] "SEA Rule 17a-4 | FINRA.org." http://www.finra.org/industry/interpretationsfor/sea-rule-17a-4. Accessed 7 Mar. 2018.



Who Benefits From Website Archiving?

Financial Services

The financial services industry is under constant pressure to adopt new online channels [5] if they are to keep up with changing digital landscape, but in doing so, their use of these channels need to be balanced against stringent compliance requirements.

This includes requirements around data retention, which is covered in MiFID II, FCA guidance, FINRA and more.

Any solution must be able to demonstrate compliance with such regulations as well as with any potential data sovereignty and GDPR requirements.

Public Sector

Numerous national archives, libraries, governments and universities now archive website data to preserve all records of cultural and historical significance. This is mostly driven by legislation such as the UK Public Records Act 1958 [6] and, more recently, the Freedom of Information Act 2000 [7].

As the public sector undertakes more activity online, organisations are looking for ways to evolve their website archiving provisions in order to take advantage of new technologies such as:

- The cloud - To allow for efficient and flexible storage of the large data sets.

- Indexing and search - To make the data useful to researchers, civil servants, students and members of the public (including public-facing portals such as The UK National Archives).

- The update from the traditional ARC file format to the ISO standard WARC file format - which can help to store born-digital or digitised materials. [8]

Brands

FMCG brands are starting to create more and more online content in addition to traditional brand assets. This can easily be altered, corrupted or lost without planning and foresight.

Keeping a record of online brand activity and customer communications can help inform future brand direction through monitoring of performance, inspire future campaigns and ensure a legally admissible record of all communications are kept for cases of dispute resolution.

[5] "Advisors ARE Social | The Putnam Social Advisor Survey 2016." https://www.putnam.com/advisor/business-building/social/. Accessed 15 Mar. 2018.

[6] "Public Records Act 1958 - Legislation.gov.uk." http://www.legislation.gov.uk/ukpga/Eliz2/6-7/51. Accessed 7 Mar. 2018.

[7] "Freedom of Information Act 2000 - Legislation.gov.uk." https://www.legislation.gov.uk/ukpga/2000/36. Accessed 7 Mar. 2018.

[8] "The WARC File Format (ISO 28500) - Information, Maintenance, Drafts." http://bibnum.bnf.fr/WARC/. Accessed 7 Mar. 2018.

About Website Archiving from MirrorWeb

The website archiving solution from MirrorWeb allows organisations to harness a powerful yet easy-to-use service that can help solve challenges around recording and retaining website data for regulatory compliance, legal and historical preservation needs.

MirrorWeb has a proven track record in helping clients across a range of sectors, including financial, public sector and even brand archivists, to meet their requirements in capturing immutable, historical records of their organisation's website.

Our solution is able to improve an organisation's operational efficiency and improve compliance with our value-added features such as:

- **Complete archives** - We archive all website content, this includes everything from internal and external sources, images, video, metadata and even social media channels.

- **Cloud-native solution** - MirrorWeb are partnered with AWS to deliver a turnkey, scalable and future-proof solution in a fully secure AWS S3 environment.

- **Close to original format** - Every archive is captured in near-real time, as it was on the day it was published.

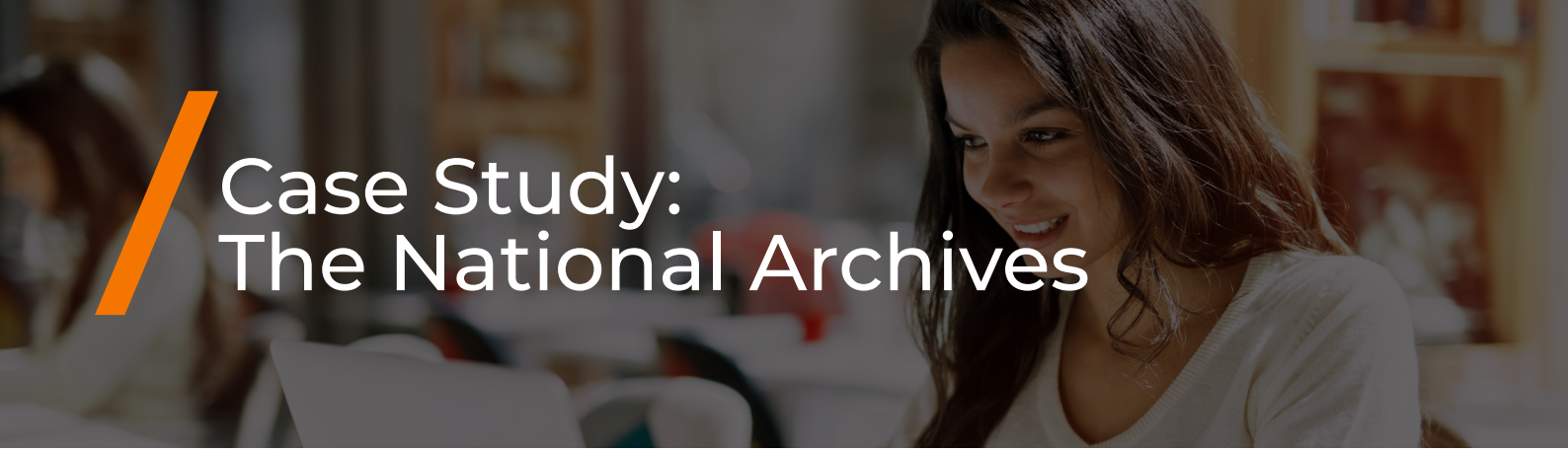
- **Full text search** - With Elasticsearch technology, all archives are indexed and searchable.

- **Sophisticated user portal** - The MirrorWeb portal allows users to be able to search and replay content, set-up and manage archive crawl frequency and parameters. The data is accessible anywhere as it is cloud-based providing flexibility and scalability with user protocols that manage functionality.

- **Public portal** - Where there is a need for public access to archives, specifically within government and national archives for example, MirrorWeb have developed a proprietary portal that integrates with the user portal to deliver high fidelity access and contemporary user experience for sharing records of cultural and historic significance with the general public.

- **Meet compliance requirements** - all archived records are stored in the ISO-standard WARC format, including date and timestamps. The MirrorWeb portal provides tools to monitor compliance risks and records can be made available to eDiscovery professionals as required.

- **Local territories** - Archives are stored in local territories being ISO9001 and ISO27001-certified and GDPR compliant.



Case Study: The National Archives

The National Archives' UK Government Website Archive is one of the largest web archiving projects in the world. Currently, the archive, which is over 150TB, contains all UK central government data published on the web from 1996 to present. This data is made up of billions of documents, web pages, images, videos and more, from all government departments.

In 2016, MirrorWeb was awarded the contract to provide the National Archives' web archiving service and set about migrating this vast amount of data from over 70 hard drives in a Paris data centre to a scalable, future-proof AWS S3 environment. Using two AWS Snowballs and two custom-built machines capable of connecting eight hard drives at once, this large-scale migration project was completed in just two weeks.

The second stage of the project was to then build a public-facing website, capable of serving over 70 million visitors a month with full replay functionality of all the archives and full text search functionality. Using Elasticsearch and a custom-built application called WarpPipe, MirrorWeb was able to index a staggering 1.4 billion documents in just ten hours - an average of 146 million documents per hour.

Today, website visitors can search the entire contents of the UK Government Website Archive - including any archived web pages and documents, with all historical links still attached as they were on the day they were published, from government departments, agencies and officials - and have at their fingertips a permanent, unaltered record of all web content published by the UK central government from 1996 to present.

About MirrorWeb

MirrorWeb delivers cloud-based archiving and monitoring solutions for the information-driven enterprise. Trusted by the UK government, our website and social media archiving platform allows organisations to create permanent, unalterable records of all online communications, meeting compliance obligations and ensuring information of commercial, cultural or historical value is never lost. Our robust web content monitoring tools also organisations to ensure content created by their representatives and partners remains compliant at all times.

The MirrorWeb platform is:

- **State-of-the-art** - offering support for web and social media data at large scale, as well as indexing for search and big data initiatives.

- **Cloud-native** - as an AWS partner, we offer near-unlimited capacity and scalability with complete control over data storage.

- **ISO-compliant** - we are ISO9001 and ISO27001-certified and archive our data in the secure, date and time-stamped ISO28500 standard WARC file (WORM) format.

- **UK-based** - we offer UK-based support 24/7/365. All archives are stored in local territories to meet data protection and compliance requirements.

- **User-friendly** - our best-in-class client portal puts you in control of your archives, allowing you to control archiving frequency, search and replay content, and view reports and notifications.

- **Cost-competitive** - we give you and your team full access to the MirrorWeb portal at all times, with no seat fee and no setup and maintenance fees.

To find out more about what MirrorWeb could do for you, get a free consultation and talk to us about your digital archiving project today.

FREE CONSULTATION

e: info@mirrorweb.com
t: 0800 222 9200

MirrorWeb

e: info@mirrorweb.com
t: 0800 222 9200
www.mirrorweb.com
Find us on