Langzeit-Webarchivierung – sicher mit PDF/A

Rechtssicherheit, Corporate Governance, serverbasierter Extraktionsdienst, Metadaten

Jeden Tag tauchen im World Wide Web tausende gefälschte Websites neu auf. Cyberkriminelle erstellen diese täuschend echt aussehenden Seiten, um Kreditkartennummern oder E-Banking-Angaben zu stehlen und so ans Geld der Nutzer zu kommen. Es gibt aber auch anders gelagerte Fälle, wie die Gerichtspraxis zeigt: So versuchten etwa Betrüger, mit einer Klage auf Basis gefälschter Webseiten einen massiv niedrigeren Preis für ein Produkt durchzudrücken. Der Betreiber der echten Website kann in einem solchen Fall den wirklich verlangten Preis am besten nachweisen, wenn er seine Webinhalte laufend archiviert.

Webarchivierung: bisher Kür statt Pflicht

Angesichts der zunehmenden Bedeutung von Firmenwebsites als Informationsquelle, von E-Commerce, Social Media und anderen webbasierten Diensten ist es fast erstaunlich, dass der Gesetzgeber bisher keine Pflicht zur Archivierung von Webinhalten kennt – obwohl Webinhalte durchaus geschäftsrelevant sein können. So publizieren Finanzinstitute im Web entscheidungsbestimmende Informationen für Investoren und sollten jederzeit in der Lage sein nachzuweisen, wann welche Angaben veröffentlicht waren.

Der Trend in der Rechtsprechung geht in den USA und der EU in die Richtung, Webinhalte wie andere Dokumente zu behandeln. Auch in der Schweiz ist juristisch nicht völlig klar, inwieweit die Website eben doch in den Einzugsbereich der Geschäftsbücherverordnung des Bundes sowie weiterer Gesetze wie dem Steuer-, Sozial- und Umweltschutzrecht fällt, die Unternehmen zur langfristigen und unverfälschbaren Aufbewahrung aller geschäftsrelevanten Informationen vom Hauptbuch bis zur Korrespondenz verpflichten.

DMS Halle 6, 6D02

www.pdf-tools.com

Nadine Schuppisser ist Leiterin Marketing bei der **PDF Tools AG** und verfügt über jahrelange Erfahrung im IT-Umfeld. Die PDF Tools AG ist ein Hersteller von Softwarelösungen und Programmierkomponenten für die PDF und PDF/A Erzeugung, Bearbeitung, Wiedergabe und Archivierung. **Daniel Spichty**, Dipl. El. Ing. ETH, ist anerkannter Spezialist für Records Management. Der Gründer und Geschäftsführer des Software-Engineering-Unternehmens The Virtual World Company hat mit seiner Firma das UBS-Archivierungsprojekt als Projektleiter unterstützt.

Langfristige Aufbewahrung bringt Vorteile

Die Vorschriften der geltenden oder künftig zu erwartenden Rechtsprechung mögen das Hauptmotiv sein, Inhalte aus dem Firmenweb zu archivieren. Doch es gibt weitere Argumente für die Web-Archivierung:

Rechtssicherheit
Wie eingangs erwähnt kann es bei einem Rechtsstreit ausschlaggebend sein, ob eine bestimmte Information zu einem bestimmten Zeitpunkt angezeigt oder eben nicht beziehungsweise nicht mehr vorhanden war. Dem lässt sich nur entgegentreten, wenn die entsprechenden Webinhalte langfristig archiviert werden und das Archiv bei jeder Anpassung der Website nachgeführt wird. In der gängigen Praxis der Website-Pflege werden die bestehenden Inhalte demgegenüber jeweils einfach durch die neue Version ersetzt – was zuvor präsentiert war, geht verloren.

PDF/A – Standard für die Langzeitarchivierung

Damit PDF-Dokumente auch für die Langzeitarchivierung gerüstet sind, wurde PDF/A geschaffen. Der Begriff bezeichnet eine Reihe von Normen, die heute unter diversen ISO-Standards etabliert sind. Für PDF/A-Dokumente gelten dabei die folgenden Maßstäbe:

- Sie sind langfristig lesbar – über Zeiträume von Jahrzehnten.
- Sie sind selbstbeschreibend – und enthalten alle Ressourcen wie Bilder, Schriftarten und Farbdefinitionen, die für eine originalgetreue Darstellung erforderlich sind.
- Sie sind eindeutig visuell reproduzierbar.
- Sie sind durchsuchbar und werden durch Metadaten beschrieben.

Es existieren verschiedene Ausprägungen von PDF/A mit teils unterschiedlichen Eigenschaften. Für die Archivierung von Webseiten genügt im Allgemeinen die lockerste Norm PDF/A-1b, welche eine visuell eindeutige Darstellung garantiert.

Corporate Governance
Neben den gesetzlichen Vorschriften existieren in allen Organisationen interne Regelungen, oft in Form von Dokumenten im Intranet. Auch diese Informationen sollten langfristig und jederzeit nachvollziehbar zur Verfügung stehen. Das gleiche gilt für öffentlich publizierte Informationen zur Unternehmensführung und Organisation wie Organigramme und Lebensläufe des Management-Teams sowie für das Impressum und die Allgemeinen Geschäftsbedingungen.

Geschichte
Gerade in KMUs ist die Unternehmensgeschichte oft am besten oder sogar einzig auf der Website umfassend dokumentiert. Neben der Firmenhistorie könnte zukünftige Betrachter zudem auch der Wandel der Website im Lauf der Zeit interessieren. Werden die bestehenden Seiten nicht archiviert, gehen Inhalte und Erscheinungsbild bei jedem Redesign unwiederbringlich verloren.

Wissen
In firmeninternen Blogs und Diskussionsforen tauschen die Mitarbeitenden Probleme und Lösungen aus. Dieser enorme Wissensschatz geht dem Unternehmen verloren, wenn bei einem Wechsel des Content-Management-Systems die bestehenden Informationen nicht archiviert werden.

Eine vollständige Archivierung aller Webinhalte kommt aus Kosten- und Kapazitätsgründen jedoch kaum in Frage. Bei der Einführung der Webarchivierung ist sorgfältig abzuwägen, welche Informationen strategisch wichtig oder anderweitig erhaltenswert sind und was problemlos „entsorgt" werden kann.

Statische oder transaktionale Speicherung

Konventionelle statische Webinhalte lassen sich durch einen serverbasierten Extraktionsdienst – meist „Grabber" oder „Crawler" genannt – für die Archivierung aufbereiten. Der Dienst prüft die zur Aufbewahrung vorgesehenen Webadressen (URLs) in regelmässigen Abständen und legt bei jeder Änderung die neue Version der Seite im Archiv ab. Üblicherweise werden dabei einfach der Code der Seite (HTML, CSS, JavaScript) sowie zugehörige Ressourcen wie Bilder und andere eingebettete Inhalte gespeichert. Für die unverfälschbare Ablage und die Integration

von Metadaten wie Zeitstempel, Stichworte und Indexinformationen ist ein nachgelagertes Archivsystem zuständig.

Bei transaktionsorientierten Webdiensten wie Online-Shops, webbasierten Produkt-Konfiguratoren und auszufüllenden Online-Formularen aller Art ist die statische Archivierung in festgelegten Intervallen ungeeignet: Hier muss der gesamte Datenstrom zwischen dem Browser des Nutzers und dem Webserver zum Zeitpunkt der Transaktion erfasst und gespeichert werden, und zwar derart, dass die Webseite später genau so rekonstruiert werden kann, wie sie der Nutzer zu Gesicht bekommen hat.

PDF/A – auch für die Ablage von Webseiten

Bei der Schweizer Großbank UBS wurde im Jahr 2009 ein bestehendes Archivsystem durch ein neues Enterprise Content Management-System abgelöst. Nachdem die schon früher archivierten Daten ins neue System übertragen waren, stellte man fest, dass auch die Unternehmenswebsite viele Informationen enthält, die ebenfalls archiviert werden müssen – sei es aus rechtlichen und geschäftspolitischen Überlegungen oder aus historischem Interesse.

Die UBS evaluierte daher auf Basis eines Anforderungskatalogs mit 15 Punkten die Web-Archivprodukte verschiedener Anbieter und stellte rasch ein essentielles Problem fest: Mit der Ablage des HTML-Quellcodes ist nicht garantiert, dass die Darstellung auch nach zehn, zwanzig oder mehr Jahren noch dem Original entspricht. Denn für die Anzeige ist nicht nur der Code verantwortlich, sondern auch die Rendering-Engine des Webbrowsers. Und die steht angesichts der rasanten Entwicklung der Webtechnologien in Zukunft mit hoher Wahrscheinlichkeit nicht mehr in der ursprünglichen Version zur Verfügung.

Die Webinhalte mussten also in einem anderen Format abgelegt werden. Als besonders geeignet wurde in diesem Zusammenhang das PDF/A-Format bewertet, das sich bereits als Standard für die Archivierung von anderen Dokumententypen etabliert hatte. Um dieser Anforderung gerecht zu werden, wurde von der PDF Tools AG der 3-Heights™ Document Converter mit einer Schnittstelle zur Web-Archivierung mit PDF/A ergänzt. Die gewählte Lösung ist vom Prinzip her einfach: Auf Basis einer manuell zusammengestellten Liste mit den URLs der zu archivierenden Webseiten arbeitet der Document Converter die UBS-Website allnächtlich ab, überprüft anhand eines einfach zu berechnenden Hash-Werts, ob eine neue Version vorliegt, generiert aus jeder geänderten Seite ein PDF/A-Dokument und übergibt dieses an das Archivsystem. Die PDF/A-Version ist im Archiv als scrollbare Seite inklusive aller Links verfügbar und lässt sich anhand der URL oder des Datums finden.

Fazit

Mit dieser Lösung können alle geschäftsrelevanten Webseiten zuverlässig archiviert werden. Das PDF/A-Dokument enthält neben der originalgetreuen Darstellung auch den indexierten Textinhalt der Webseite – Volltextsuche ist also ebenfalls unterstützt.