# Transaktionales Archivieren



Version 1.01

*Die Archive der Welt sehen sich zunehmend gezwungen, Web-Inhalte zu archivieren. Mit der Entwicklung des Web wird diese Aufgabe immer schwieriger, da die ehemaligen „statischen“ Websites heute zunehmend von „dynamisch“ zusammengesetzten bei jedem einzelnen Abruf für jeden einzelnen Benutzer speziell konfektioniert zusammengestellt und niemals mehr auf dieselbe Art vorkommen werden. Diese Tatsache erfordert eine neue klare Positionierung und Definition der Aufgaben eines Archivs. Stattdessen haben einige Archive als neues Paradigma das „transaktionale Archvieren“ eingeführt[1]. Statt dem Publizierten soll das Abgeholte, also der Verkehr auf der Leitung zwischen Client und Server, archiviert werden[2].*
*Damit erhöht sich die Nachvollziehbarkeit der vom Server tatsächlich an den Client kommunizierten Inhalte. Auch ist man beim Archivieren des Verkehrs auf der Web-Schnittstelle von all den komplizierten Fragen nach der visuellen Präsentation der Web-Inhalte befreit. Schliesslich ist bekannt, dass über 90% der CMS-Seiten einer Verwaltung oder der Bestände einer Bibliothek oder eines Archivs überhaupt nie abgerufen werden und folglich gemäss der transaktionalen Logik gar nicht archiviert zu werden brauchen.*
*Was wie die eierlegende Wollmilchsau für Archive aussieht, entpuppt sich auf den zweiten Blick als Überbietung aller Überwachungswunschträume aller Diktatoren und Geheimdienste im Osten und im Westen welche im harmlosen Deckmäntelchen der Erhaltung des kulturellen Erbes daherkommt.*

## *Die fundamentale Verwirrung: Das Internet ist nicht homogen!*

Schon das Konzept einer „Webarchivierung“ basiert auf der darunterliegenden Vorstellung Internet-ferner Menschen, dass „das Web“ ein homogenes Medium – etwa wie ein modernes Fernsehen – sei. Dieser Irrtum kommt vielleicht daher, dass wir mit Anderen im Internet immer über rechteckige Browserfenster kommunizieren. Wenn „das Web“ zuerst über Nachbildungen des Telefons – etwa mit den heutigen Smartphones – erschlossen worden wäre, wäre nie ein Archivverantwortlicher auf die Idee gekommen, sämtliche telephonische Kommunikation zu archivieren. Solche Ideen erwarten wir eher von der Polizei oder vom Geheimdienst. Selbst wenn man sich auf das Archivieren der Kommunikation mit Amtsstellen beschränkt, müsste man bei jedem Gespräch

1 http://ictk.ch/content/nachvollziehbarkeit-und-compliance-bei-webinhalten
2 http://www.dpconline.org/component/docman/doc_download/865-dpctw13-01pdf

einen Hinweis plazieren, dass es aufgenommen wird und wozu die Aufnahme verwendet wird.

Tatsächlich ist das Internet so hochgradig heterogen wie die sonstige Kommunikationswelt um uns herum. Es gibt intime, private Ecken, es gibt Broadcast-Medien, es tummeln sich dort Vereine, politische Parteien, Firmen, Bekannte, Freunde und Kriminelle wie im wirklichen Leben. Da sich digitale Information hervorragend eignet, um jede traditionelle Kommunikationsform nachzustellen und viele neue zu erfinden, verdrängt die Welt der digitalen Kommunikation zusehends alle „analogen“ Kommunikationsmedien.

Die Idee, „das Web“ zu archivieren, ist also vergleichbar mit der Vorstellung, „die Welt der Kommunikation“ zu archivieren. Das ist definitiv nicht die Aufgabe eines Archivs!

## *Die erste Verwirrung: Cache, Protokoll, Sicherheitskopie, Historisierung, Versionierung und Archiv*

Wenn man die Werbung der Firmen liest, die transaktionales Archivieren von Webinhalten anbieten, so trifft man auf ein heilloses Durcheinander von Konzepten, die alle etwas mit Speichern zu tun haben. Es lohnt sich, diese Konzepte einzeln zu betrachten und voneinander abzugrenzen. Denn obwohl alle diese Konzepte irgendwie technisch teilweise durch Nachbarkonzepte ersetzt werden können, sind doch der Zweck und das legale Umfeld sehr unterschiedlich. Es ist hochgradig problematisch, wenn ein Archiv sich bemüht, die Aufgabe einer Cache, eines Behälters von Logdateien, einer Backup-Station oder einer historisierten Datenbank zu übernehmen. Denn deren Randbedingungen und Ziele sind völlig verschieden von denen eines Archivs.

### Cache

Mit Cache bezeichneten die Polarforscher Vorratshütten mit Proviant, die sie auf dem Hinweg anlegten, um sie auf dem Rückweg zu nutzen. Das anglisierte Wort für Versteck (vor Eisbären) wird wohl immer noch mit einem französischen „ch“ ausgesprochen, weil diese Erfindung aus Kanada stammt.

Im Kontext des Internet bezeichnet Cache ein Zwischenlager von Daten, welches es überflüssig macht, den ganzen Weg zur Datenquelle zu gehen. So kann etwa der Browser eine Website lokal in einer Cache speichern und beim nächsten Abruf dem Benutzer schnell präsentieren, sofern sich keine Inhalte beim Server verändert haben. Auch beim Server können häufig gelesene Seiten im Hauptspeicher gecachet werden, damit sie nicht jedes Mal erneut als Datei geladen werden müssen. Und zwischen Server und Browser können noch beliebig viele Zwischenstationen eine Cache anlegen, um die Antwortzeiten tief zu halten. Der Zweck einer Cache ist nicht dass Archivieren, sondern nur die kurzfristige Zwischenlagerung zum Zweck der Optimierung der Antwortzeiten. Eine Cache wird laufend bedenkenlos gelöscht, da sie ja immer nur aus Kopien von anderweitig angebotenen Daten besteht.

Die Daten-Inhalte einer Cache gehören immer Anderen. Ein Cache-Betreiber ritzt Urheberrechte und die Privatsphäre und hat die Pflicht, sämtliche Inhalte einer Cache äusserst vertraulich zu behandeln und sie oft zu löschen.

Oft wird nun die Eignung des transaktionalen Archivierens als Cache-Mechanismus gepriesen. Wenn transaktionales Archivieren auf einem Webserver eingerichtet ist, wird eine Webseite nicht bei jeder Änderung abgespeichert, sondern nur, wenn eine Webseite abgerufen wird, die sich gegenüber der letzten „archivierten“ („archiviert“, ist das oft tendenziös verwendete Synonym für „abgespeichert“) Version verändert hat. Das ist sicher ein hübscher Cache-Mechanismus. Er wird aber der Aufgabe des Archivars nicht gerecht. Wenn wir alle Gesetze abschaffen würden, die nie in einem konkreten Prozess „abgerufen“ werden, müssten wir mehr als 90% aller Gesetze abschaffen. Nur schon das Wissen um die Existenz von Gesetzen verändert aber unser Verhalten. Nur schon das Wissen um behördliche Publikationen hat eine Auswirkung, auch wenn diese nie abgerufen werden.

## Protokoll

Das Wort Protokoll wird hier im Sinne von Sitzungsprotokoll verwendet und auf englisch jeweils als Log bezeichnet. (In anderen Zusammenhängen steht es für Kommunikationsregeln und wird auf englisch als Protocol bezeichnet.) Protokolldateien werden von fast jedem Programm geschrieben. Sie zeichnen oft im Detail auf, was in einem Programm ablief. Sie dienen den Programmierern zum Verfolgen von Abläufen und Aufsuchen von Programmierfehlern. Sie dienen den Betreibern zur Überwachung als Behälter von Warnungs- und Fehlermeldungen, die eine Einschränkung des geordneten Funktionierens eines Programms verzeichnen. Im Zusammenhang mit „dem Web“ verzeichnen sie manchmal alle kommunizierten Inhalte, oft aber auch nur Verbindungsdaten („Randdaten“), welche angeben, wer wann mit wem kommuniziert hat. In dieser Funktion sind Log-Dateien meistens nur kurzfristig nützlich und sollten nach wenigen Tagen oder Wochen vom Betrieb gelöscht werden.

Solche Log-Dateien sind neuerdings ein von Polizei und Geheimdiensten heiss umkämpftes Feld, welche diese Datenquelle für zeitnahe, aber auch jahrelang aufbewahrte, Basis für Rasterfahndung verwenden möchten. Diesen Wünschen stehen diverse gesetzliche Schranken im Weg, welche die Privatsphäre schützen. Als Material für die eigentliche Archivierung eignen sich Protokolldateien wenig, denn traditionelle Archive haben zum Zweck, die Nachvollziehbarkeit der Arbeit von Behörden zu gewährleisten und deren Tätigkeit zu dokumentieren. Die Überwachung privater Kommunikationen darf erst dann Archivgut werden, wenn der Geheimdienst seine Dossiers beim Archiv abgibt.

## Sicherheitskopie

Eine Sicherheitskopie dient wie der Name schon sagt, der Sicherung von Daten und soll vor Datenverlust schützen. Sie ist naturgemäss nur kurzfristig gültig und wird meistens periodisch mit aktuelleren Daten überschrieben. Ihre langfristige Aufbewahrung ist nicht geplant. Oft besteht sie aus seltener angelegten Vollkopien und häufiger erzeugten Differenzkopien. Ein Datenkontext ist für die Sicherheitskopie nur implizit im Arbeitsablauf gegeben. Das Dateneigentum liegt nur teilweise bei der Institution, welche Sicherheitskopien herstellt.

Auch eine Sicherheitskopie steht im potentiellen Konflikt mit Urheberrechten und mit dem Schutz der Privatsphäre. Wer sie anlegt, ist verpflichtet, ihre Inhalte genauso vertraulich zu behandeln wie die Originaldaten. Wer keinen Zu-

gang zu den Originaldaten hat, darf ihn nicht auf dem Umweg über die Sicherheitskopie erschleichen.

Grundsätzlich eignet sich die Technik des transaktionalen Archivierens schlecht als Sicherheitskopie schlecht, da Daten, die noch nicht abgerufen wurden, nicht gesichert werden.

## Historisierung

Unter einer „historisierenden“ Datenbank versteht man eine Datenbank, in welcher nachvollzogen werden kann, wie der jetzige Zustand der Datenbank erreicht wurde. Es werden also alle Änderungen an der Datenbank mit Zeitstempel und Benutzeridentifikation verzeichnet. Theoretisch kann man den Zustand der Datenbank zu einem beliebigen früheren Zeitpunkt aus den Daten der Historisierung rekonstruieren.

Oft werden auch nur einzelne Tabellen einer Datenbank „historisiert“. So will man etwa in einer Bank überprüfen können, wer wann wie ein Kundenkonto manipuliert hat, wer das Verzeichnis von Postleitzahlen alle paar Jahre auf den neuen Stand bringt, interessiert dagegen die interne Compliance-Abteilung nicht.

Die Technik des transaktionalen Archivierens hat eine gewisse Ähnlichkeit mit der Historisierung. Beim ersten Abrufen einer veränderten Seite wird diese mit Zeitstempel und Abrufdaten gespeichert. Allerdings ist es hier der Lesevorgang, der die Historisierung auslöst, nicht der Schreibvorgang.

Bezüglich Privatsphäre und Urheberrecht stellen sich für historisierte Datenbanken genau dieselben Probleme wie für alle Datenbanken. Urheberrechtlich sind die einzelnen Inhalte von Datenbanken oft mangels Schöpfungshöhe nicht geschützt. In vielen Ländern ist die Datenbank in ihrer Gesamtheit urheberrechtlich geschütztes Eigentum. Der Schutz der Privatsphäre ist juristisch besonders bei Datenbanken ausgeprägt, wobei oft vergessen geht, dass ein Dateisystem, eine statische Website oder ein Content Management System (CMS) auch nur Datenbanken sind.

## Versionierung

Schon seit dem Dateisystem auf alten DEC-Maschinen kennt man die Versionierung von Dateien. Um dem Datenverlust vorzubeugen, werden Dateien nie gelöscht, sondern alle Versionen des Dateisystems sind jederzeit verfügbar. Oft wird bei jedem Einfügevorgang in das Versionssystem nur die Differenz der Dateien zur Vorversion gespeichert. Insofern wächst die Datenmenge im Versionssystem bedeutend langsamer als man auf den ersten Blick annehmen würde.

Fast alle Programmierer verwenden Versionsverwaltungen für ihre Arbeit. Bei Nichtprogrammierern setzt sich das Konzept erst langsam wieder durch. Neben dem Schutz vor Datenverlust erleichtert die Versionierung auch die gemeinsame Arbeit an einem Dokument ungemein, da dieses nicht mehr im Kreis herumgeschickt werden muss, sondern an Ort und Stelle von allen Mitarbeitern geändert werden kann, ohne dass ein Chaos entsteht. Neuere Geschäftsverwaltungssysteme von Behörden speichern mehrere Versionen eines Dokuments oder verwenden die Historisierungsmechanismen der zugrundeliegenden Datenbank, um die Nachvollziehbarkeit und Zuschreibbarkeit von Änderungen zu gewährleisten.

Versioniert wird grundsätzlich alles, was ins Versionssystem eingespeichert wird. Naturgemäss sind das allerdings vor allem Daten, deren Dateneigner die Benutzer sind.

Das System der transaktionalen Archivierung ist besonders stolz auf seine Versionierung, die nicht vom Urheber, sondern vom Abrufenden ausgelöst wird. Nur geschriebene und nie gelesene Versionen existieren in diesem Archiv nicht. Allerdings fragt man sich, wie man etwas schreibt, ohne es gleichzeitig zu lesen. Die CD-WOM (Write-Only Memory) ist immer noch nicht erfunden ...

### Archiv

Ein Archiv unterscheidet sich fundamental von allen anderen aufgezählten Methoden der Ablage und des Abspeicherns. Schon in vordigitaler Zeit gab es neben den Archiven viele andere Methoden der Ablage und der Abspeicherung, die den oben aufgezählten digitalen Techniken ähneln: Ablagen, Notizem, Laufwege, Dossiers, ... Trotzdem wäre niemand je auf die Idee gekommen, die Ablage der Zeitungsausgaben der letzten Woche am Kaffeetisch einer Redaktion als Archiv zu bezeichnen.

Archive haben zum Zweck, die Aktivität einer Institution nachvollziehbar langfristig zu dokumentieren. Der Auftrag behördlicher Archive ist oft sogar gesetzlich geregelt. Im Bundesgesetz über die Archivierung[3] heisst es

> [1] Rechtlich, politisch, wirtschaftlich, historisch, sozial oder kulturell wertvolle Unterlagen des Bundes werden archiviert.
>
> [2] Die Archivierung leistet einen Beitrag zur Rechtssicherheit sowie zur kontinuierlichen und rationellen Verwaltungsführung. Sie schafft insbesondere Voraussetzungen für die historische und sozialwissenschaftliche Forschung.

Hier ist festzuhalten, dass nur Unterlagen der Institution zu archivieren sind. Das sind von der Institution selbst hergestellte Unterlagen oder solche, welche ihr von Dritten absichtlich zum Zweck der Bearbeitung und Archivierung überlassen wurden. Das Archiv erfüllt also weder die Aufgaben von Cache oder Log, noch von Sicherheitskopie oder Versionsverwaltung, denn es dient weder zur Überwachung des Verhaltens der Bürger noch zur Erleichterung der täglichen Arbeit der Organisationseinheiten der Institution.

## *Die zweite Verwirrung: Datenbanken und Websites*

Oben wurde angemerkt, dass Websites auch Datenbanken sind. Im heute üblichen Sprachgebrauch sind aber Datenbanken oft mit Datenschutzregeln belegte Sammlungen von Information einer Institution, welche von ihren Mitarbeitern beim Ausführen ihrer Aufgabe benutzt werden. Solche Datenbanken sind im allgemeinen nicht öffentlich, sondern sogar intern vielfältig gegen unbefugte Zugriffe geschützt.

---

3 http://www.admin.ch/opc/de/classified-compilation/19994756/index.html

Websites dagegen sind öffentlich und gleichen somit einer Publikation der Institution.

Wenn wir traditionelle Archivmuster heranziehen, entspricht eine Datenbank der Kollektion von Dossiers einer Behörde, während eine Website dem Publikationsmedium (Prospekt, Merkblatt, Zeitschrift, Buch) entspricht.

Die Dossiers einer Behörde werden von dieser vertraulich behandelt und erst mehrere Jahre nachdem ein Dossier geschlossen wurde, wird es dem Archiv übergeben. Auch dort kann in den ersten 30 oder 50 Jahre nur diejenige Behörde darauf zugreifen, welche das Dossier der Behörde übergeben hat. Ein Dossier kann Daten – etwa Zuschriften, Anträge, etc. - von Dritten enthalten. Das geistige Eigentum an diesen Daten ist nicht an die Behörde übergegangen. Da Dossiers oft personenbezogene Daten enthalten, müssen die Behörde und das Archiv die jeweils für diesen Datenbestand geltenden Regeln des Datenschutzes einhalten.

Die Publikationen einer Behörde hingegen sind oft nicht vom Urheberrecht geschützt.

## *Die dritte Verwirrung: Publizieren und Konsumieren*

Bei einer Website, die man mit dem URL (Universaler Ressourcen-Lokator) einer Behörde abruft, scheinen alle Inhalte von dieser Behörde zu stammen. Dies ist im allgemeinen nicht wahr. Vielmehr ist ein beträchtlicher Teil dessen, was im Browserfenster erscheint, vom abrufenden Benutzer und nicht von der Behörde erschaffen. Der Benutzer legt Farbe und Grösse der Schrift fest. Wenn er ein Formular ausfüllt, wird ihm das Resultat seiner Eingaben vom URL der Behörde übermittelt angezeigt. Dass trotzdem er und nicht die Behörde deren Urheber und Dateneigner ist, ersieht er schon daran, dass er sich überall registrieren und mit seinem Passwort anmelden muss, damit er nicht unzulässigerweise Zutritt zu den Daten Dritter erhält. Alle individuell auf den Benutzer zugeschnittenen Elemente einer Seite sind streng genommen nicht von der Behörde erzeugt, sondern vom Benutzer verantwortet. (Das gilt natürlich auch für nicht behördliche Institutionen und ihre Archive, auch wenn deren Ziel nicht gesetzlich sondern bloss statutarisch festgelegt ist.)

Oft wird in heutigen Websites unter einem URL eine sogenannt „dynamische“ Seite publiziert. Das heisst, dass der URL nicht auf den dargestellten Inhalt schliessen lässt. Vielmehr werden mit einer unter dem Namen AJAX bekannten Technik je nach Benutzerinteraktion Daten im Hintergrund von anderen unsichtbaren URLs heruntergeladen und als Dokumentfragmente in die dargestellte Seite eingefügt. Am Ende entsteht eine nur für diesen Benutzer nur zum jetzigen Zeitpunkt zusammengestellte Browserseite, deren Urheber grösstenteils der Benutzer selber ist, während allenfalls gewisse Textversatzstücke aus dem Content Management System (CMS) der Behörde stammen.

Insofern ist das Archivziel, die Browsersicht des Benutzers zu archivieren nicht nur unerfüllbar, sondern widerspricht auch dem Auftrag des Archivs. Dieses hat nicht den Auftrag, die Benutzer und seine genaue Interaktionssequenz zu überwachen, sondern die (exekutive) Aktivität der Behörde langfristig zu dokumentieren. Der Zustand des Browsers des Benutzers ist ebenso wenig zu archivieren, wie in der analogen Zeit der Hersteller der Brille, die Beleuchtung oder die Perspektive eines Briefs der Behörde, wenn er vom Bürger gelesen wurde, noch Ort und Zeit dieses Geschehens. Nicht einmal die Tatsa-

che des Lesens einer offiziellen Publikation ist zu archivieren. Denn bei Gesetzen, Verordnungen und behördlichen Mitteilungen schützt Unkenntnis vor Strafe nicht. Im Gegenzug benötigt der Staat keine Information über den Kenntnisstand seiner Bürger. Wäre das Internet zuerst auf dem Telefon genutzt worden und nicht auf dem Bildschirm, würde niemand auf die Idee kommen, die Inhalte sämtliches Telefonate seien unter Angabe der Gesprächteilnehmer, deren Aufenthaltsorte und der Zeitspanne zu archivieren.

Ein Archiv hat grundsätzlich nicht den Informationskonsum der Kommunikationspartner einer Institution aufzubewahren, sondern deren Informationsproduktion. Das transaktionale Archivieren verkehrt diesen Grundsatz in sein Gegenteil und eignet sich deshalb hervorragend zur Überwachung der Bürger, die sich heute kaum einem Internet-Kontakt mit Behörden entziehen können.

## *Wer nutzt transaktionales Archvieren?*

Heute gibt es erst wenige Implementationen des transaktionalen Archivierens. Diese werden vornehmlich von CMS-Systemen wie MediaWiki, Drupal, Wordpress eingebunden und daher wohl schon von vielen kleineren Website-Betreibern genutzt.

Grössere Archive diskutieren diese neue Technik, die ihnen sehr attraktiv erscheint, da sie das Problem der „dynamischen“ Websites zu lösen scheint und dem Auftrag der Website-Archivierung besser gerecht wird als ein Ernte-Roboter der jeweils auf die Reise geschickt wird, um den aktuellen Stand des Web abzuholen.

Was diesen Archiven besonders attraktiv in der durch Abfragen ausgelösten Archivierung erscheint, ist die Tatsache, dass sich der Archivroboter nirgends mit Passwort anzumelden braucht und die Archivierung der Inhalte auf Seiten des Servers die Verschlüsselung umgeht. So erhält der Archivierungsmechanismus zentralen Einblick in alle vertraulichen Kommunikationen mit der Behörde, ohne selber seine Zugriffsrechte beweisen zu müssen.

## *Was ist die Aufgabe eines Archivs?*

Wir kommen also zum Schluss, dass die Aufgabe eines Archivs nicht das „Archivieren des Web“ sein kann. Die Aufgabe des Webarchiving ist von vornherein falsch gestellt und muss von den verantwortlichen Archivdirektoren, den Regierungen, den Parlamenten und dem Volk korrigiert werden, damit die Erfüllung dieses Auftrags nicht zur Hydra der totalen Überwachung mutiert.

Wenn ein Bürger über eine „e-Governement“-Plattform (aus einem unerfindlichen Grund wird solchen Online-Aktivitäten oft ein kleines „e-“, ein „i“ oder Ähnliches zur Geisterbeschwörung vorangestellt) benutzt, sollen seine Aktivitäten in ein Dossier in einer vertraulich verwalteten Datenbank der betreffenden Behörde fliessen und keinesfalls dann schon archiviert werden. Zehn Jahre nachdem das betreffende Dossier geschlossen wurde, kann die Datenbank dem Archiv mit den für den betreffenden Dossiertyp angemessenen Datenschutzauflagen zur Langzeitaufbewahrung übergeben werden. Eine „Webarchivierung“ dieser Inhalte ist fehl am Platz.

Wenn eine Institution Webseiten oder Webseitenfragmente (etwa in einem CMS) publiziert, gibt es einen Arbeitsablauf der Publikation, in welchem zu einem klar definierten Zeitpunkt eine Freigabe der Inhalte durch dazu autori-

sierte Personen erfolgt. Der Moment der Freigabe ist der Punkt, wo die Archivierung der behördlichen Informationen die Inhalte sinnvollerweise abzweigt. Dazu muss das Archiv nicht wissen, wem eine Seite wie dargestellt wird, sondern nur, wer welche Inhalte wann im CMS publiziert hat.

## *Müssen die Archive den Geheimdiensten unterstellt werden?*

Es gibt viele Berührungspunkte zwischen Geheimdiensten und Archiven. Aus Sicht der Geheimdienste wünschen sich diese, dass die Archive ihnen möglichst flächendeckend Daten liefern, die der vertraulichen Kommunikation zwischen der Behörde entstammen und die ohne Passwörter, Aufknacken der Verschlüsselung oder Staatstrojaner nur schwer erhältlich sind.

In der umgekehrten Richtung vom Geheimdienst zum Archiv fliessen die Daten nomalerweise eher spärlich. Die Archive könnten sich den ganzen Aufwand mit der Webarchivierung sparen, wenn sie einfach die bei der NSA zu den Behörden-URLs gehörigen Daten erhalten könnten.

Wenn die transaktionale Archivierung in den Archiven eingeführt wird, sollte man von vornherein auf die harmlose Front der Archive verzichten und sie gleich ganz in den Geheimdienst eingliedern. Dann wäre wenigstens klar, welchen Gesetzen sie unterstellt sind.