

Zeitschrift: Arbido
Herausgeber: Verein Schweizerischer Archivarinnen und Archivare; Bibliothek Information Schweiz
Band: 14 (1999)
Heft: 5

Artikel: "On the dark side of the cyberspace" : zur Archivierung des Internets
Autor: Hagmann, Jürg
DOI: <https://doi.org/10.5169/seals-769098>

Nutzungsbedingungen

Die ETH-Bibliothek ist die Anbieterin der digitalisierten Zeitschriften. Sie besitzt keine Urheberrechte an den Zeitschriften und ist nicht verantwortlich für deren Inhalte. Die Rechte liegen in der Regel bei den Herausgebern beziehungsweise den externen Rechteinhabern. [Siehe Rechtliche Hinweise.](#)

Conditions d'utilisation

L'ETH Library est le fournisseur des revues numérisées. Elle ne détient aucun droit d'auteur sur les revues et n'est pas responsable de leur contenu. En règle générale, les droits sont détenus par les éditeurs ou les détenteurs de droits externes. [Voir Informations légales.](#)

Terms of use

The ETH Library is the provider of the digitised journals. It does not own any copyrights to the journals and is not responsible for their content. The rights usually lie with the publishers or the external rights holders. [See Legal notice.](#)

Download PDF: 01.04.2025

ETH-Bibliothek Zürich, E-Periodica, <https://www.e-periodica.ch>

«ON THE DARK SIDE OF THE CYBERSPACE»

ZUR ARCHIVIERUNG DES INTERNETS

von Jürg Hagmann

Die grosse Alexandrinische Bibliothek (*Museion*) mit rund 700 000 Buchrollen ging 47 v. Chr. im Krieg mit Cäsar grösstenteils in Flammen auf. Frühe Buchdruckerzeugnisse sind nicht erhalten geblieben, und viele Filme aus der Pionierzeit mussten aufgrund des Silbergehalts entsorgt werden. Die Entstehungsgeschichte jedes neuen Mediums ist eine Geschichte des Verlusts und ihrer partiellen Rekonstruktion durch Fragmente.

ERHALTUNG DER DIGITALEN GESCHICHTE

Soll dieses Schicksal auch die Gründungsphase des Internets ereilen? Eine mutige Gruppe von Unternehmern und IT-Spezialisten um den Kalifornier *Brewster Kahle* (San Francisco) hat sich vorgenommen, dass die Anfangsphase der Digitalgeschichte erhalten werden soll.

Ist dies nicht ein völlig aussichtsloses Unterfangen? Wie können zukünftige Historiker von einer im besten Fall bruchstückhaften Erhaltung des Web profitieren, das täglich um rund 1,5 Millionen Seiten wächst, wie Spezialisten schätzen, wobei die durchschnittliche Lebensdauer einer Webseite nur gerade 75 Tage beträgt¹.

Obwohl also das Internet ein äusserst flüchtiges Medium ist, hat es sich ein Mann in den Kopf gesetzt, die vergänglichen Inhalte für die Nachwelt zu konservieren. Nach Abschluss seiner Ausbildung am MIT 1982 baute Kahle in den achtziger Jahren bei der Firma *Thinking Machines* Super-Parallelrechner mit Tausenden von Prozessoren. Anfang der neunziger Jahre entwickelte er WAIS, eine der besten Suchtechniken für das Netz und Namensgeber für sein erstes Unternehmen, das er 1995 für 15 Mio \$ an den Online-Dienst AOL verkaufte.

Für Kenner der Online-Szene war es denn auch keine Überraschung, dass gerade Kahle einer der ersten war, die Antwort auf eine drängende Frage suchten: Wie lässt sich das Netz archivieren? «Für digitale Informationen gibt es bisher kein Äquivalent zu einer Bibliothek», schrieb er im März 1997 im Wissenschaftsmagazin *Scientific American*². Wenn sich daran nichts ändere, drohe der Menschheit ein unwiederbringlicher Verlust. Im folgenden wird die Botenschaft aus diesem wegweisenden Artikel zusammengefasst.

ZIELSETZUNG DES INTERNET-ARCHIVS

Im Vordergrund von Kahles Projekt steht also die Erhaltung des kulturellen Erbes im digitalen Zeitalter. Nie zuvor in der Geschichte der Menschheit gab es ein so reichhaltiges «kulturelles Artefakt» wie das Internet, auf das so einfach zugegriffen werden konnte und das der globalen Lehre und

Forschung dient. «Unsere Technik soll helfen, dass Menschen besser und schneller kommunizieren können. Das macht Gemeinschaften intelligenter.»³

Der bildungspolitische Idealismus droht jedoch am flüchtigen Medium bzw. der Web-Infrastruktur selbst zu scheitern:

- Das Internet ist unzuverlässig und eignet sich nicht für das wissenschaftliche Zitieren. Zu oft kommt beim Surfen die Meldung «404 document not found». Der Web-Experte *Jakob Nielsen* hat in einem Bericht⁴ diese ins Nichts führenden Links als ernsthafte Gefahr für das Web bezeichnet. Allerdings gibt es auch eine Website, die sich nur auf das Auffinden von 404-Meldungen spezialisiert hat (*www.cool404.com*)¹. Man könnte dahinter eine interessante Kultur des Scheiterns von Webseiten entdecken.
- Der gefundenen Information mangelt es oft an Kontext und Authentizität: Wo befinde ich mich? Kann ich dieser Information vertrauen?
- Es gibt keine weiteren Findmittel, wenn ich in einer Sackgasse lande (wohin navigieren?). In der Bibliothekswelt gibt es immerhin noch ansprechbare BibliothekarInnen sowie einen interbibliothekarischen Leihverkehr.

PROBLEME DER BESTANDESERHALTUNG

Um die Gefahr der verschwindenden Informationen abzuwenden, starteten Kahle und einige Freunde 1997 in San Francisco das *Internet-Archiv* (*www.archive.org*). Ihr Ziel: Das Netz so komplett wie möglich abzuspeichern, also nicht nur sämtliche öffentlich zugänglichen Web-Seiten, sondern auch die *Gophers* und die *Netnews* sowie *download-fähige Software*.

Problem der Haltbarkeit digitaler Information: Während bei älteren Papierdokumenten (industriell gefertigte Holzschnittpapiere) der Zerfall droht, erweist sich die EDV für die Langzeitarchivierung von Information bisher immer noch als ungeeignet. Kein Lieferant gibt eine Garantie ab, wie lange CDs, Disketten oder optische Platten lesbar sind, zudem ist mit erheblichen Kosten zu rechnen, wenn die Daten nach Ablauf der Lebensdauer des Datenträgers

¹ Vgl. Weber, Daniel: Jäger der verlorenen Websites. Immer mehr verschwindet im digitalen Orkus, in: NZZ 19.2.1999, S.65

² vgl. <http://www.sciam.com/0397issue/0397kahle.html> oder www.archive.org (daselbst ein Link zum Volltext-Artikel)

³ vgl. Link-Dossier in der Wochenzeitung «Die Zeit» v. 12.3.1999 zum Thema Internet-Archiv:

<http://www2.bdaserver.de/zeit/tag/link-dossier/Computer/Internet/Archiv/>

⁴ vgl. www.useit.com/alertbox/980614.html

migriert werden müssen. Kommt noch hinzu, dass auch die entsprechende Hard- und Software archiviert werden muss, wenn maschinenlesbare Daten aufbewahrt werden. Das Internet-Archiv von Kahle plant einen Migrationsrhythmus von 10 Jahren für Daten und Betriebssysteme. Damit ist jedoch nur der weniger anspruchsvolle Teil des Problems gelöst.

Die Konversionsdynamik der verschiedenen File-Formate (Text, Bild, Audio und Video) ist kaum in den Griff zu bekommen. Kahles Team denkt nur daran, in die Archivierung der bekannten und verbreitetsten Formate zu investieren. Weitere Investitionen müssen in die Datensicherheit sowie in die Migration von *Metadaten* (Informationen über die gespeicherten Informationen) gemacht werden.

«Wie wollen Sie ohne solche Metadaten herausfinden, ob Sie etwa einem Online-Reisebüro trauen können oder nicht?»⁵ Um unfreundlichen Übernahmeangeboten durch Dritte vorzubeugen, die einen exklusiven kommerziell motivierten Zugriff auf ein solches Archiv anstreben könnten, übertrug Kahle die Datenrechte auf eine gemeinnützige Organisation, die über genügend Mittel verfügt, um die nötige Maintenance der Speicherung über Jahre zu garantieren.

TECHNISCHE ASPEKTE / MENGengerüst

Der Aufbau des Internet-Archivs bedingt einen grossen organisatorischen Aufwand analog der Bewirtschaftung eines riesigen elektronischen Archivs. Kahle könnte sich bei der Sammlung, Speicherung und Bewirtschaftung von Terabytes an verschiedenen Projekten im Bereich «*Electronic Records Management*»⁶ orientieren; der *Internationale Archivrat (ICA)* hat zudem 1996 ein entsprechendes Manual verabschiedet⁷. Von welchem Mengengerüst ist auszugehen, wenn wir alle Subsysteme des Internets einbeziehen: *WWW, Gopher, FTP und Netnews*? Kahle schätzt den öffentlich zugänglichen Teil des Internets auf ca. 2 Terabytes. Zum Vergleich: der Inhalt aller Bücher der *Library of Congress* (rund 20 Mio. Bände) in Washington würde etwa 20 Terabytes umfassen. Bei einem Wachstum von über 1,5 Mio. Seiten pro Tag käme das Netz im Jahr 2000 auf über eine Milliarde Seiten (1 Seite im Web entspricht durchschnittlich 30 KB inkl. Grafiken).

Mengen (1997)

	Websites	Total	Wachstum
WWW	> 400000	2 TB	600 GB/Monat
Gopher	5000	100 GB	abnehmend
FTP	10000	5000 GB	unbekannt
Netnews	20000	240 GB	16 GB/Monat

(1 Gigabyte = 1000 Megabytes, 1 Terabyte = 1000 Gigabytes. 1 GB reicht, um 1000 Bücher zu speichern oder 1 Std. komprimiertes Videoformat.)

Um die im Internet-Archiv abgelegte Information wieder zu finden, hat Kahle eine eigene Suchmaschine entwickelt, die zugleich den neuen Firmennamen abgibt: *Alexa*⁸. Ende 1998 hat *Alexa* der *Library of Congress* zwei Terabyte Web-Inhalt geschenkt, den «*Snapshot*» des WWW-Jahrgangs 1997.

Im Moment kann es sich Kahle noch leisten, praktisch alle Daten aus dem öffentlich zugänglichen Teil des Web zu sammeln (d.h. exkl. Inhalte, wo der Zugang durch Passwörter bzw. *subscriptions* eingeschränkt ist). Doch der Moment wird kommen, wo eine Selektion bzw. Bewertung der Daten stattfinden muss, wie auch Kahle einräumt.

KOSTENASPEKTE

Bei diesen Datenmengen ist sicher die Frage nach der kostengünstigsten Speichertechnologie eine der zentralsten. Rund 200 Dollar kostet es, ein Gigabyte auf Festplatte abzuspeichern. *Tapes* mit *Jukeboxen* schlagen dagegen nur mit zwanzig Dollar pro Gigabyte zu Buche.

Kostenübersicht Speichertechnologien (Zugriffszeiten)

RAM	12000 \$/GB (70 Nanosek.)
Hard Disk	200 \$/GB (15 Millisek.)
Optical Disk Jukebox	140 \$/GB (10 Sek.)
Tape Jukebox	20 \$/GB (4 Min.)
Tapes vom Gestell	2 \$/GB (manuelle Unterstützung)

Das Internet-Archiv arbeitet mit Hard Disks für häufig benutzte Daten und mit Tape-Jukeboxen für die grosse Masse.

RECHTLICHE UND SOZIALE ASPEKTE

Für sein umfassendes Angebot erntet Kahle allerdings nicht nur Lob. Kritische Geister wenden ein, dass sein Dienst Urheberrechte verletze und vor allem wenig Rücksicht auf den Datenschutz nehme: So mancher Netzbürger wolle nun einmal nicht mehr wahrhaben, was er einst auf seiner persönlichen Homepage von sich gegeben habe – während in Alexas Datenbanken auch die Jugendsünden noch alle nachzulesen sind. Kahle ficht diese Kritik nicht an. «Unsere Regel ist klar», erwidert er. «Wer sich in seinem Urheberrecht verletzt fühlt oder Informationen aus der Welt schaffen will, kann uns anweisen, die entsprechenden Dateien zu löschen. Aber bisher haben das nur sehr wenige getan. Im Gegenteil – sie wollen Teil der digitalen Geschichte sein.»⁹

NEUE DIENSTE UND PROFITABILITÄT

Kahle ging es von Anfang an nicht nur darum, die Datenmassen sicher abzulegen. Er wollte sie auch nutzbar

⁵ Ludwig Siegele stellt Brewster Kahle aus San Francisco vor, der das scheinbar Unmögliche möglich macht. Er archiviert die gigantischen Datenmengen des Internet und hofft, damit Geld verdienen zu können. Vgl. Link-Dossier in der Wochenzeitung «Die Zeit» v. 12.3.1999; URL vgl.³⁾

⁶ vgl. die Projekte der Universitäten Pittsburgh, Edith Cowan und British Columbia, siehe: *Electronic Records Management, A literature review* by the International Council of Archives (Ottawa 1996, S.22–29) oder online <http://www.archives.ca/ica/archives/site1/p-er/english.html>

⁷ Guide for managing electronic records from an archival perspective, ICA 1996, <http://www.archives.ca/ica/archives/site1/p-er/english.html>

⁸ <http://www.alexa.com>

⁹ <http://www.zeit.de/links>

machen. Deswegen gründete er 1998 die Firma Alexa, benannt in Anlehnung an die Bibliothek von Alexandria, die bedeutendste Dokumentensammlung der Antike. Mit seinen Diensten im Bereich *Textmining* betreibt Kahle eine Art Wissensbewahrung im Dienste eines globalen Knowledge-Managements. Im Vordergrund steht dabei ein «*Reliability service*» für Dokumente, die vom ursprünglichen Urheber nicht mehr erhältlich sind. Ein wichtiger Schritt zu einer Infrastruktur, in der das weltweite Hypertext-System als Medium des wissenschaftlichen Zitierens anerkannt wird. Ein wichtiges Angebot von Alexa besteht darin, Netznutzern mit intelligenten Navigationsvorschlägen zu helfen, jene Web-Dienste im elektronischen Weltgespinnst zu finden, die sie auch wirklich interessieren. Dazu merken sich die Alexa-Rechner, welche virtuellen Wege Web-Surfer gehen. Kahle: «Wenn Sie auf einem Trampelpfad durch den Wald gehen, dann profitieren Sie von den Exkursionen Ihrer Vorgänger. Genau das Prinzip wenden wir auf das Netz an.»

Wer sich von Alexa durch das Netz führen lassen will, muss sich ein kleines, kostenloses Programm vom Web-Dienst des Unternehmens herunterladen. Es erweitert die Browser-Software von *Netscape* oder *Microsoft* um eine Menüleiste am unteren Rand des Bildschirms. Darauf klicken die Nutzer, wenn sie die Surf-Empfehlungen sehen wollen. Dort finden sich freilich noch andere interessante Buttons – etwa jenen, der ein Fenster mit Informationen über den gerade besuchten Dienst öffnet: Wer betreibt den Service? Wie beliebt ist er? Wievielen anderen Web-Angeboten ist er einen Link wert, einen elektronischen Querverweis?

«Das eine grosse Problem im Internet ist, Informationen zu finden», begründet Kahle diese Fülle von Angaben, «die andere, mindestens ebenso grosse Schwierigkeit ist, dass man nie genau weiss, wo man landet». Richtig ernst nimmt der Alexa-Chef dagegen die Zweifel der Experten, ob er mit seinem Angebot jemals Geld verdienen könne. Wie bei den meisten anderen Web-Diensten auch, soll vor allem Werbung für die nötigen Einnahmen von Alexa sorgen. «Als einzige Geldquelle reicht das auf die Dauer nicht», meint *Jerry Michalski*, Chefredakteur des renommierten Branchenbriefes *Release 1.0*.

Kritische Stimmen wie etwa der Experte für Geschichte und elektronische Artefakte, *Edward Higgs*,¹⁰ halten wenig bis gar nichts von Archiven im Cyberspace. Higgs moniert dabei v.a. die eingeschränkte «Tunnelsicht» einer auf Cyberarchive beschränkten Historiographie.

Als Archivar kann ich dem beipflichten, denn ohne Kontext von Provenienzen und Geschäftsprozessen (*Verwaltungszusammenhänge*, s. *ARBIDO 4/99*, *M. Schaffroth*) kann kein konsistentes Verständnis für historische Sachverhalte hergestellt werden. Dazu kommt noch, dass der heuristische Prozess in einem Archiv selbst viel Input (*cross-fertilization*) durch menschliche Kommunikation liefert.

Higgs: «*The very fact being in an archive, talking to archivists and other readers in a quasi-social setting, is important for such cross-fertilization.*»¹¹

Dazu kommt auch folgender Umstand: «*The really interesting discoveries in research are often made through happy accident whilst browsing through archival material, or when insights from one research field are applied to another.*»¹²

Dieses unter dem dokumentarischen Begriff «serendipity»¹³ bekannte Phänomen könnte allerdings auch auf das *Retrieval* in elektronischen Dokumenten angewendet werden.

Solche Einwände halten den Internet-Archivar aber nicht davon ab, über die technische Zukunft seines Dienstes nachzudenken. Sein Ziel: eine Art virtuelle Zeitmaschine. Nutzer sollen bei ihm abfragen können, wie ein Web-Angebot zu einem bestimmten Zeitpunkt ausgesehen hat. So will er seine Technik immer weiter verbessern. Einmal, so hofft Kahle, wird sie den Menschen sogar helfen, die schweren Fragen des Lebens zu beantworten: Welches Buch soll ich lesen? Soll ich auf die Universität gehen? Wie soll ich meine Kinder erziehen? «Ich will», träumt er, «ein richtiges Orakel von Delphi schaffen.»¹⁴

⁹ <http://www.zeit.de/links>

¹⁰ Higgs hat als Herausgeber des Buchs «History and electronic artefacts», (Oxford 1998), anlässlich eines internationalen Kolloquiums die aktuelle Diskussion zusammengefasst.

¹¹ Vgl. *Electronic Records Management*, ICA 1996, S.74

¹² ebenda

¹³ Das zufällige Finden von relevanten Informationen, nach denen man aber ursprünglich gar nicht gesucht hat (insbesondere bei Online-Recherchen); man spricht auch von «unpredictable cross-sense».

¹⁴ <http://www2.bdaserver.de/zeit/tag/link-dossier/Computer/Internet/Archiv/>

Information und Dokumentation Fachhochschul-Diplomstudium

Knowledge-Management
– Chance ergreifen

Wir sind für Sie da und helfen Ihnen gerne weiter!

HTA

Chur

Hochschule
für Technik+Architektur
Ringstrasse
7004 Chur

Tel 081 286 24 24

Fax 081 286 24 00

sekretariat@fh-htachur.ch

www.fh-htachur.ch