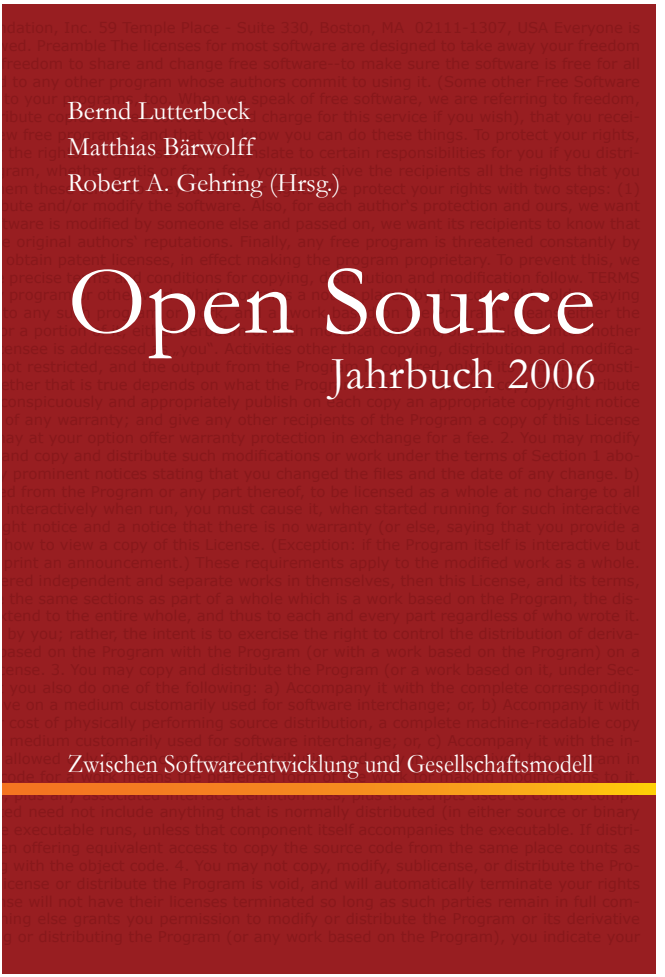
Dieser Artikel ist Teil des

Open Source Jahrbuchs 2006



erhältlich unter www.opensourcejahrbuch.de.

Die komplette Ausgabe enthält viele weitere interessante Artikel. Lob und Kritik zu diesem Artikel sowie weitere Anregungen können Sie uns einfach und unkompliziert mitteilen per E-Mail oder auf www.opensourcejahrbuch.de/feedback/.

Ein Archiv für die ganze Welt

STEFFAN HEUER

Brewster Kahle ist einer der führenden Köpfe der Open-Content- und Open–Access-Bewegung. Sein 1996 gestartetes *Internet Archive* ist das Gedächtnis des Webs mit 55 Milliarden gespeicherten Webseiten. Sie lassen sich auch Jahre nach ihrem Verschwinden aus dem Netz noch im Originalzustand aufrufen. Der Computerwissenschaftler will so die Bücherei von Alexandria wieder auferstehen lassen – als ein offenes Archiv von Texten, Bildern, Ton- und Videodateien. Das wirft erhebliche technische wie rechtliche Probleme auf. Seit kurzem haben sich Hochschulen, Bibliotheken und namhafte Technologiefirmen seiner *Open Content Alliance* angeschlossen. Parallel dazu ringen die Suchmaschine Google, Medienkonzerne und Verlage mit ihren eigenen, umstrittenen Visionen des Zugangs zu digitalen Werken.

Schlüsselwörter: Internet Archive · Langzeitarchivierung · Online-Bücherei · Open Content Alliance · Urheberrecht

1 Das Internet Archive und die Wayback Machine – von der Jugendvision zum Web-Archiv mit 55 Milliarden Einträgen

Als Student am *Massachusetts Institute of Technology* in Cambridge hatte Brewster Kahle vor 25 Jahren eine Vision: Computer-Netzwerke würden Leben und Arbeit der Informationsgesellschaft grundsätzlich verändern. Kahle kam beim Nachdenken über die vernetzte Zukunft auf zwei Dinge, denen er seine beruflichen Anstrengungen widmen würde. Entweder wollte er dafür sorgen, erinnert er sich rückblickend, dass private Daten sicher verschlüsselt werden oder er wollte eine Online-Version der legendären Bücherei von Alexandria entwickeln. „Seitdem ist mir keine neue Idee eingefallen. Und daraus ist eine glänzende Karriere geworden“, sagt der Computerwissenschaftler.

Entschieden hat er sich für letztere Mission. Nachdem er mehrere IT-Unternehmen gewinnbringend verkaufte, widmet sich Kahle seit einem Jahrzehnt seinem Motto: der

Welt „universellen Zugang zum menschlichen Wissen" zu verschaffen. Dazu gründete Kahle 1996 das *Internet Archive* in San Francisco.[1] Das Herzstück des virtuellen Archivs befindet sich in einem unscheinbaren Holzhaus aus dem 19. Jahrhundert im ehemaligen Militärgelände Presidio in Sichtweite der Golden-Gate-Brücke. Von hier aus dirigiert Kahle mit 30 Mitarbeitern das bislang ehrgeizigste Projekt, einen Großteil des menschlichen Kulturerbes in Text, Bild, Ton und Video zu katalogisieren und über das Web zugänglich zu machen. Das als Non-Profit betriebene Archiv wird von privaten Spenden finanziert – darunter mehrere Millionen aus Kahles Privatvermögen.

Bislang hat seine Stiftung rund 26 000 Bücher und andere Texte erfasst, die entweder nicht mehr dem Urheberrechtsschutz unterliegen oder im Rahmen einer Open-Content-Lizenz frei herunterladbar sind. Das gleiche gilt für annähernd 29 000 Videoclips, Fernsehserien und historische Filme sowie rund 69 000 Audiodateien und 30 000 Konzertmitschnitte. Das Herzstück der Sammlung allerdings sind rund 55 Milliarden Webseiten, von denen viele nicht mehr online zugänglich sind, da sie vom Netz genommen oder aktualisiert wurden.[2]

Die *Wayback Machine* ist, wie der Name suggeriert, eine Zeitmaschine, die das wohlbekannte Problem der Fehlermeldung „404 – Page not found" löst. Benannt hat sie Kahle nach der Zeitmaschine aus der Zeichentrickserie „Rocky und Bullwinkle". Die Reise in die Online-Vergangenheit ermöglichen Programme oder Software-Bots, die Webseiten sammeln und wie eine gigantische Sammlung von Schnappschüssen archivieren. Über eine Suchmaske kann ein Nutzer eine Web-Adresse eingeben und das angezeigte Ergebnis bis auf den Tag genau eingrenzen. Die *Wayback Machine* archiviert das Web in aller seiner Vielfalt – von Nachrichten und politischen Ereignissen oder Debatten, über den Internetauftritt von Unternehmen, Hochschulen und Forschungseinrichtungen bis zu den Seiten über Alltag und Hobbys von Privatleuten lässt sich so wieder aufrufen, was anderweitig längst im Informationsorkan der modernen Welt untergegangen ist.[3]

Gehostet werden die Seiten in bislang drei ehemaligen Lagerhäusern im Stadtteil *South of Market* in der Innenstadt von San Francisco sowie auf Servern in Amsterdam und Alexandria, dem Standort seines großen Vorbildes. Kahles visionäre Sammelwut hat in einem knappen Jahrzehnt einen Datenberg von gut einem Petabyte[4] angehäuft. Jeden Monat, schätzt Kahle, wächst seine Sammlung um rund 20 Terabyte – was mehr als der gesamte Bestand der *Library of Congress* ist.[5]

1 Das *Internet Archive* findet man unter http://www.archive.org. Eine weitere hervorragende Einführung ist ein Interview, das Kahle 2004 gab. Dieses kann man unter: http://www.itconversations.com/shows/detail400.html herunterladen.

2 Die Statistiken wurden mit Stand vom 10. Februar 2006 auf http://www.archive.org abgerufen.

3 Siehe http://www.archive.org/web/web.php.

4 Ein Petabyte sind 1024 Terabyte oder circa eine Million Gigabyte.

5 Siehe http://www.archive.org/web/hardware.php.

2 Das Ausgangsproblem: Wie viele Informationen gibt es und wie lassen sie sich am besten erfassen und archivieren?

Die Datenflut einer vernetzten Welt zu erfassen und anschließend gratis zur Verfügung zu stellen, wirft erhebliche Probleme auf – archivarischer, informationstechnischer, finanzieller und schließlich rechtlicher Art.

Da ist zunächst einmal die Frage, welche Werke genau ein offenes Internet-Archiv erfassen soll. Nie zuvor hat die Menschheit derart viele Bits und Bytes produziert. Peter Lyman von der Schule für Informations-Management und -Systeme an der Universität Berkeley versucht seit 2000 mit einer oft zitierten Studie namens „How Much Information?", das Datenaufkommen zu quantifizieren (Lyman und Varian 2003). „Seit Mitte der 90er Jahre hat sich etwas Erstaunliches getan", berichtet der Wissenschaftler. „Der Zugriff auf Informationen wurde dank des Internets immer besser und schneller, während die Kosten für Speichermedien immer weiter sinken." Beide Trends erlaubten es Büchereien, Museen, Stiftungen, Unternehmen und Regierungsbehörden in aller Welt, immer mehr Daten zu horten und nach Brauchbarem zu durchsieben.

Im Jahr 2002 etwa produzierte die Menschheit auf Papier, Film, magnetischen und optischen Speichermedien rund fünf Exabyte neue Daten. Neun Zehntel davon wurden auf Festplatten abgelegt. Fünf Exabyte – eine Zahl mit 18 Nullen – entspricht 800 Megabyte pro Kopf der Weltbevölkerung, rechnet Lyman vor, oder zehn Meter aneinander gereihter Buchrücken pro Erdbewohner. Seit 1999, dem Zeitpunkt der ersten Erhebung für die Berkeley-Studie, wuchs das Volumen gespeicherter Informationen um jährlich 30 Prozent. Das für alle zugängliche und nicht hinter Passwörtern und Firewalls geschützte Internet macht dabei nur einen Bruchteil aus, nämlich 170 Terabyte. „Das ist nur die Oberfläche – statische Seiten, die für jeden Benutzer gleich aussehen", erklärt der Professor. Rechnet man all jene Seiten hinzu, die aufgrund einer individuellen Anfrage dynamisch erzeugt werden, erhält man das so genannte „dunkle Web" – und das ist geschätzte 92 000 Terabyte (oder knapp 90 Petabyte) groß.

Ähnliche Kalkulationen hat Kahle für Kinofilme, Schallplatten und CDs angestellt, aufbauend auf Schätzungen des Computerwissenschaftlers Raj Reddi von der Carnegie Mellon Universität. Reddi schätzt, dass in der gesamten Menschheitsgeschichte seit Zeiten der sumerischen Keilschrift rund 100 Millionen Bücher veröffentlicht wurden. Grafisch originalgetreue digitale Versionen würden allerdings weitaus mehr als 1 MB pro Buch erfordern. Dazu kommen zwei bis drei Millionen Musikaufnahmen vom 78er Format über LPs bis zu CDs, rund 100 000 Kinofilme sowie weitere zwei Millionen Filme, die zu Bildungs-, Werbe- oder anderen Zwecken aufgenommen wurden. Selbst die jährliche Ausbeute aller in den USA ausgestrahlten Fernsehsendungen beläuft sich auf technisch handhabbare 3,6 Millionen Stunden, die nach Lymans Schätzungen maximal 8 200 Terabyte Speicherplatz benötigen. Suchmaschinen-Weltmeister Google rechnete unlängst den Zeitaufwand durch, um alle auf der Welt vorhandenen

Informationen mit seiner Technik zu indexieren und kam auf 300 Jahre.[6]

Kahles Sammeleifer geht längst nicht so weit. Er konzentriert sich in erster Linie auf frei zugängliche Webseiten. Und da kämpft das *Archive* gegen die Uhr. Die durchschnittliche Lebensdauer einer Webseite beträgt 77 Tage, die von offiziellen Seiten ungefähr vier Monate. So schnell, wie sie aufgeschaltet werden, verschwinden Dokumente wieder oder können innerhalb von Minuten verändert werden, sollten Fehler entdeckt werden oder Beschwerden eingehen. Automatische Archivierungs-Software macht nur sporadisch die Runde durchs Web und bringt eine Momentaufnahme zurück, die bereits veraltet ist, wenn sie online verfügbar ist. So dauert es nach Angaben des *Internet Archive* mindestens sechs, in der Regel zwölf Monate, bis eine indexierte Seite über die *Wayback Machine* verfügbar ist.[7]

Die Wirtschaftswelt hat den Wert der *Wayback Machine* bereits erkannt. Durchschnittlich 180 000 Nutzer suchen nach archivierten Internet-Dokumenten, und immer häufiger sind es Rechtsanwälte oder Unternehmer auf der Jagd nach Beweismaterialien. Das *Wall Street Journal* etwa beschrieb im Sommer 2005, wie Anwälte des Computerherstellers Dell Kahles Service benutzten, um alte Versionen einer Webseite zu finden, die ihre Marke angeblich diffamierte und Besucher zu anderen PC-Anbietern umleitete. Der Betreiber hatte die reklamierte Seite längst gegen eine harmlosere Version ausgetauscht, aber bei der *Wayback Machine* wurden sie fündig und erreichten, dass die beanstandete Domain an Dell übertragen wurde. Laut *Wall Street Journal* ist Kahles Archiv bereits in den Sprachgebrauch übergegangen. „Can you do a Wayback on that?" erkundigen sich Juristen bei Kollegen, wenn sie heute eine Stecknadel im Online-Heuhaufen suchen (Kesmodel 2005).

3 Keine rein technische Frage: Sind Software-Bots die besten Archivare?

Die *Wayback Machine* ist ein automatisierter Archivar, dessen Roboterprogramme indexieren, was sie finden – so lange sie es finden. Das klingt auf den ersten Blick wie wertneutrale Technik und ähnelt der Herangehensweise kommerzieller Suchmaschinen wie Google, Yahoo! oder A9.com, die ihrerseits Milliarden von Webseiten durchkämmen und im Cache verfügbar machen. In fast allen Fällen gilt, dass sich selbst mit einer schnellen Suche finden lässt, wogegen der Eigentümer oder ein Betroffener nicht nachträglich Einwände erhebt. Aber Automatisierung ist längst keine problemfreie Lösung. „In traditionellen Archiven treffen Menschen Entscheidungen, welche Werke aufbewahrt werden sollen. Im Internet ist das eine mathematische Entscheidung, deren Algorithmen Firmen wie Google als Geschäftsgeheimnis hüten: Alle

6 Die Schätzung stammt von Google-CEO Eric Schmidt auf der Jahreskonferenz der *Association of National Advertisers*. Siehe auch Mills (2005).

7 Detaillierte Informationen zu Lebensdauer und Verzögerung bis zum Erscheinen im *Archive* sind unter http://www.archive.org/about/faqs.php zu finden.

soundsoviel Tage zieht ein Roboter Bilanz", sagt Lyman. Das mag objektiver sein als ein Kurator am Königshof oder ein Beamter in einem totalitären Staat, der unliebsame Bestände aus dem Verkehr ziehen kann – eine umfassende Abbildung der Wirklichkeit ist es jedoch ebenso wenig.

Weiterhin stellt sich die Frage, wie weit ein Archiv das Netz auswerten sollte. Jede Webseite verweist im Schnitt auf 15 andere Seiten und enthält 5 Objekte wie Bilder, Grafiken, Videos, Tondateien, Werbung.[8] Wer eine Seite archiviert, muss beim zu speichernden Umfeld irgendwo eine Grenze ziehen und droht damit wichtigen Kontext zu kappen, der der Nachwelt das Verständnis erleichtert. Ähnlich ergeht es Forschern, die heute alte Zeitungen lesen, die es nur noch auf Mikrofilm gibt. Annoncen, unterschiedliche Lokalausgaben oder Korrekturen nach dem Andruck sind oft unwiederbringlich verloren. Wenn die Links zudem auf Quellen verweisen, die im „tiefen Web" verborgen sind, kommt der normale Nutzer nicht mehr weiter. Viele solcher Fundstücke sind oft nur in kostenpflichtigen Datenbanken – etwa denen von Tageszeitungen oder Lexis-Nexis – erhältlich.

Neben der richtigen Suche gibt es das Problem der richtigen Aufbewahrung. Selbst wenn sich heute jede Seite durch Bots archivieren ließe, heißt das noch lange nicht, dass zukünftige Generationen in der Lage sein werden, diese Dokumente aufzurufen. Browser etwa gibt es erst seit rund zehn Jahren – und schon jetzt schwankt die Darstellung erheblich, je nachdem, ob man eine Seite mit einer alten Version von *Netscape*, Microsofts *Internet Explorer* oder Apples *Safari* aufruft. Die Hardware und Software, um ein Web-Objekt authentisch darzustellen, muss bewahrt werden oder kompatibel sein, sonst drohen die gehorteten Informationen wertlos zu sein.[9]

Dazu gehören die Metadaten – also etwa Informationen über Herkunft und Authentizität der Quelle. Sonst kann es passieren, dass spätere Forscher Webseiten voller Verschwörungstheorien über die Anschläge vom 11. September finden und sie mit offiziellen Berichten der verschiedenen Untersuchungskommissionen und seriösen Analysen gleichsetzen. Ganz zu schweigen von der Gefahr, dass sich veraltete Formate überhaupt nicht mehr öffnen lassen, weil sie inkompatibel sind oder sich die Speichermedien zersetzt haben, auf die der Server zugreifen will. Das *Internet Archive* speichert die Webseiten gegenwärtig in Dateien von je 100 Megabyte Größe im ARC-Format, das Kahle bereits 1996 entwickelte. Die wichtigsten Kriterien dabei sind die Autonomie der in diesen Paketen gespeicherten Dateien, dass sie sich also ohne Index-File finden und öffnen lassen. Das Format muss „ausbaubar" sein, so dass die

8 Die Angaben stammen aus einem Essay Peter Lymans für das NDIIP-Projekt. Siehe http://www.digitalpreservation.gov/index.php?nav=3&subnav=5.

9 Grundsätzliche Fragen zur richtigen, dezentralen Aufbewahrung öffentlicher digitaler Bestände versucht auch das „Electronic Records Archives"-Projekt (ERA) zu beantworten. Dazu gehören neben rund fünf Milliarden Seiten in Washington auch die Bestände aus dreizehn regionalen Archiven und bislang elf Präsidenten-Bibliotheken. Die Infrastruktur baut Lockheed Martin im Auftrag der *National Archives and Records Administration (NARA)* in den kommenden sechs Jahren für 308 Mio. Dollar auf (Squeo 2005).

Dateien mit unterschiedlichsten Protokollen aufgerufen werden können. Die Dateien müssen sich ferner streamen lassen, und sie müssen als eigenständige Einträge ohne späteres Inhaltsverzeichnis Bestand haben.[10]

Über die Frage nach der besten Aufbewahrung denkt seit einigen Jahren auch der US-Kongress nach. Berkeley-Professor Lyman und Kahle sind Berater eines 2000 in Washington lancierten Projektes, um ein Nationales Programm für Digitale Infrastruktur und Aufbewahrung (*NDIIPP*) zu entwickeln. Unter der Leitung der *Library of Congress* arbeiten US-Universitäten an Machbarkeitsstudien zum richtigen Umgang mit Materialien, die „digital zur Welt gekommen" sind. Die ersten Förderprogramme mit einem Volumen von insgesamt 17 Millionen Dollar vergab Washington im September 2004 und Mai 2005. Sie reichen von einem Auftrag, neue Archivierungswerkzeuge für öffentliche Datenbanken zu entwickeln, über die Speicherung digitaler Fernsehprogramme bis zur Aufbewahrung von „gefährdeten" Materialien aus der schnelllebigen Dotcom-Ära. Wenn eine der Universitäten eine Suite von Software–Werkzeugen fertig stellt, sollen andere Büchereien und Sammlungen sie verwenden können, und zwar möglichst als kostenlose Open-Source-Programme.[11]

4 Gewöhnliche Linux-PCs ermöglichen den Zugang für hunderttausende Nutzer

Die dritte große technische Frage stellt sich beim Zugang für alle potentiellen Nutzer. Kahles Organisation jongliert die Nutzlast auf mehreren hundert PCs, auf denen das Open-Source-Betriebssystem Linux läuft. Finanziell ist diese Form der Archivierung durchaus machbar. „Nehmen wir nur den Bestand der *Library of Congress*: rund 26 Millionen Objekte", rechnet Brewster Kahle vor. „Der reine Text in einem Buch sind rund ein Megabyte, also geht es um 26 Terabyte. Das lässt sich auf Linux-Servern für 60 000 Dollar zugänglich machen und der Welt zum Stöbern anbieten. Wollte man die Bücher scannen und grafisch aufbereiten, würde es zehn Dollar pro Band kosten. Also beliefe sich die gesamte Rechnung auf 260 Millionen Dollar. Das sind Peanuts!"

Jeder seiner Rechner verfügt über 512 Megabyte Arbeitsspeicher und eine Festplatte mit einem Gigabyte Kapazität. Daneben werden Daten auf Bändern gespeichert. Die langfristige Finanzierung der Infrastruktur und des Zugangs zu den Beständen sieht Kahle als gesichert an. „Wir werden die Bestände in eine Reihe von Archiven rund um die Welt verteilen", erklärt er zur Zukunft des Archivs. Bei dieser Anstrengung kommt ihm seine frühere Karriere als Computerwissenschaftler und erfolgreicher IT-Unternehmer zugute. So gründete und verkaufte er seine Firma *WAIS* (*Wide Area Information Systems*) und die Suchmaschine *Alexa Internet* gewinnbringend. Letztere steuert heute noch die von Bots besuchten Seiten zum *Internet Archive* bei.

10 Siehe http://pages.alexa.com/company/arcformat.html.

11 Eingehende Informationen zu Zielsetzung, Zeitplan und Pilotprojekten finden sich unter http://www.digitalpreservation.gov.

Langfristig ist die Speicherung schnell vergänglicher Artefakte wie Webseiten eine lohnende Investition, glaubt Kahle. Internet-Büchereien können demnach den Inhalt des Netzes zum permanenten Fundus unseres politischen und kulturellen Lebens machen. Sie schützen außerdem das Recht der Bevölkerung auf umfassende Information durch Behörden und Politiker. So gibt es etwa in den Vereinigten Staaten keine landesweit einheitliche Handhabe, wie und für welchen Zeitraum öffentliche Dokumente online aufbewahrt werden. „Es gibt nur wenige Regelungen, was genau aufgeschaltet wird, wann es wieder verschwindet und wie oft es aktualisiert wird. Diese Lücke können Online-Büchereien schließen", heißt es auf der Webseite des Internet-Archivs.[12]

Ebenso wichtig ist es, dass die Bevölkerung im digitalen Zeitalter ein „Recht auf Erinnerung" hat, insbesondere dann, wenn der Zugriff auf gedruckte Unterlagen nicht möglich ist, da mehr und mehr Dokumente nur noch in digitaler Form existieren. Der Archivar des Internets listet eine ganze Reihe weiterer Beweggründe auf: die Veränderung der Sprache und die Evolution des Webs als Spiegelbild der Gesellschaft und Volkswirtschaft zu verfolgen; und schließlich das Internet und seine ständig wachsenden Verbindungen von Seite zu Seite als Einblick in die Kommunikationsmuster der Menschheit zu bewahren, damit sie künftige Wissenschaftler studieren können.

5 Das Internet Archive als kulturelles Gedächtnis und seine Vorgänger

Es gab bereits frühere, durchaus umfangreiche Online-Kollektionen, die Texte, Bilder, Ton- und Filmdateien seit Jahren sammeln, sofern diese nicht (oder nicht mehr) dem Urheberrecht unterliegen. Mit ihnen arbeitet das *Internet Archive* zusammen – etwa dem Projekt *Gutenberg*, das der Amerikaner Michael Hart bereits 1971 an der Universität Illinois startete und das bislang 17 000 Bücher umfasst. Seine Vision nannte er „Replikator-Technik" – was einmal in einen Computer eingegeben wurde, lässt sich endlos vervielfältigen. Rund zwei Millionen Exemplare im guten alten ASCII-Format laden sich Nutzer in aller Welt jeden Monat herunter. Darunter sind Klassiker wie die Notizbücher des Leonardo da Vinci oder Mark Twains „Abenteuer des Huckleberry Finn".[13]

Ein weiteres wichtiges Online-Archiv, das die Verbindung zwischen gedruckter und digitaler Welt herstellt, ist das *Million Book Project*. Diese Bücherei wurde von fünf Akademikern an der Carnegie Mellon Universität in Pittsburgh gestartet und hatte ursprünglich vor, bis Ende 2005 eine Million Bücher einzuscannen, mit optischer Buchstabenerkennung (OCR) zu verarbeiten und im Internet zur Verfügung zu stellen.

12 Einer der Versuche, verschollene öffentliche Webseiten von US-Behörden zu archivieren, ist der *Cyber Cemetary*: http://govinfo.library.unt.edu/.

13 Siehe http://www.gutenberg.org. Nicht zu verwechseln mit dem deutschen Gutenberg-Projekt, das von *Spiegel Online* betrieben wird und einen ähnlichen Ansatz verfolgt (http://gutenberg.spiegel.de/).

Ein erster Schritt sind mehr als 10 500 Bände, die ebenfalls über das Portal des *Internet Archive* abgerufen werden können. Darunter sind bislang eher obskure Bücher wie „Eine Geschichte der Hindu-Zivilisation unter britischer Herrschaft, Band 1".[14]

Der dritte bedeutende Fundus ist das Video-Archiv des New Yorker Verlegers Rick Prelinger. Er begann 1983 in seiner Wohnung in Manhattan damit, längst vergessene Werbe- und Fortbildungsfilme amerikanischer Unternehmen und Interessensverbände sowie ausrangierte Kurzfilme zu katalogisieren. Heute umfasst sein Archiv rund 48 000 solcher Filme oder Bruchstücke aus den Jahren 1927 bis 1987. Die Sammlung wurde 2002 von der *Library of Congress* erworben, die Verwertungsrechte liegen bei der Fotoagentur Getty Images. Nur 2 000 dieser Videodateien sind über das *Internet Archive* zugänglich.[15]

6 Der Rechtsstreit um die Ausweitung des Urheberrechts

Prelinger und Kahle wurden vor allem durch ein Verfahren vor dem Obersten Gerichtshof der USA bekannt, in dem sie 2004 gegen die beständige Ausweitung des Urheberrechtsschutzes klagten. Während Copyright bis 1976 angemeldet und regelmäßig verlängert werden musste, geht das Gesetz in den USA heute von einer automatischen Geltung aus. Andernfalls muss ein Rechtsinhaber (Verlag, Autor) sein *opt-out* geltend machen. Da das in der Regel nicht geschieht, behalten Millionen von Titeln automatisch ihr Copyright und werden zu so genannten verwaisten Werken (*orphaned works*). Prelinger und Kahle argumentierten, wenn der Kongress regelmäßig dagegen stimme, Filme oder fiktive Charaktere wie Mickey Maus in die „Public Domain" übergehen zu lassen, verstoße dies gegen das in der US-Verfassung verbriefte Recht auf freie Meinungsäußerung.[16]

Ihr prominentester Fürsprecher ist der Jurist Lawrence Lessig von der Universität Stanford. Er ist mit mehreren Büchern und engagierten Aufsätzen zum Thema Copyright eine Galionsfigur der Open-Source-Bewegung geworden und Mitbegründer von Creative Commons. Der Oberste Gerichtshof in Washington wies die Klage Ende 2004 erwartungsgemäß ab, so dass der Fall jetzt vor dem 9. Bezirksgericht in San Francisco verhandelt wird. Es geht dabei um eine fundamentale Frage des Zugriffs auf online verfügbare Archivmaterialien – egal ob sie „digital geboren" wurden oder erst nachträglich in Bits umgewandelt wurden. Diese Debatte hat sich inzwischen zu einem weitaus größeren Problem entwickelt als die technischen Hürden der umfassenden Archivierung. Denn selbst ein Schnappschuss im Web unterliegt streng genommen

14 Das Projekt ist inzwischen Bestandteil der *Universal Library*: http://tera-3.ul.cs.cmu.edu.

15 Interessante Details zu Prelingers Motivation und der Entwicklung seines Archivs finden sich in einem persönlichen Essay des Gründers (Prelinger 2001).

16 Hintergrund zu den Eingaben Prelingers und Kahles sowie Klageschriften und Regierungsentgegnungen im Verfahren Kahle gegen Ashcroft und später Kahle gegen Gonzales finden sich auf der Webseite des *Stanford Center for Internet and Society*: http://cyberlaw.stanford.edu/about/cases/kahle_v_ashcroft.shtml.

dem Urheberrecht und darf nur dann archiviert werden, wenn der Rechteinhaber ausfindig gemacht und sein Einverständnis eingeholt werden kann.

7 Die Open Content Alliance – kommerzielle Interessen stoßen hinzu

Der wichtigste Meilenstein in Kahles Plan ist die gemeinsam mit Yahoo! im Oktober 2005 ins Leben gerufene *Open Content Alliance (OCA)*. Das Bündnis ist eine Zusammenarbeit des *Internet Archive* mit großen, traditionellen Sammlungen sowie einigen der bekanntesten Namen der Technologiewelt: Microsoft, Adobe, Hewlett–Packard Labs und der Technikverlag O'Reilly Media. Auf akademischer Seite gehören der Organisation renommierte Universitäten wie Columbia, Emory oder Johns Hopkins, die zehn Hochschulen des University-of-California-Systems und die Universität Toronto an. Dazu kommen Einrichtungen wie das Europa-Archiv unter Leitung des schweizerischen Bundesarchivs, die *Smithsonian Institution*[17] und das britische Nationalarchiv. Alleine zwischen dem Start des Projektes Anfang Oktober bis Mitte Dezember 2005 wuchs die Zahl der Teilnehmer von 10 auf 33 an.

Die *OCA* hat sich zum Ziel gesetzt, „ein permanentes Archiv digitalisierter Texte und Multimedia-Inhalte in vielen Sprachen aufzubauen", heißt es auf der Webseite der Organisation. Die Sammlungen, die Teilnehmer zur Digitalisierung bereitstellen, wird über die OCA-Seite verfügbar sein und zudem über die Suchmaschine von Yahoo!. Alle Inhalte sollen frei und kostenlos zugängig und zum Download verfügbar sein – sofern die Inhaber der Urheberrechte ihre Zustimmung geben. Den Anfang wird eine Sammlung von 18 000 Titeln klassischer amerikanischer und internationaler Literatur machen, die sich im Besitz der Universität von Kalifornien befinden.[18]

Das Digitalisieren wird unter der Leitung von Kahles Archiv erfolgen, der dafür Ende Oktober einen eigens entwickelten Scanner namens *Scribe* vorstellte. Ähnliche Geräte sind seit geraumer Zeit an großen Universitätsbüchereien wie Stanford und Toronto bereits im Einsatz. Allerdings kosten vollautomatische Systeme wie die der Schweizer Firma 4Digital Books, die bis zu 3 000 Seiten die Stunde schaffen, rund eine halbe Million Dollar.[19]

Kahles mobiler Scanner erinnert an eine tragbare Dunkelkammer mit zwei Digitalkameras. Bücher werden in eine V-förmige Glasvorrichtung eingespannt, das Umblättern besorgt allerdings ein Techniker, der auch die Aufnahme mit Pedalen steuert. Die Digitalisierung einer Seite kostet zehn US-Cent, und die gesamte Software, um eine

17 Die *Smithsonian Institution* ist ein Museumskomplex bestehend aus 19 Museen und Galerien – hauptsächlich in Washington, D.C. angesiedelt. Sie verwaltet über 142 Mio. Artefakte.

18 Die Zielsetzung und eine aktuelle Liste der Teilnehmer findet sich unter http://www.opencontentalliance.org.

19 Bilder des Scribe-Scanners von der Open-Library-Eröffnungsveranstaltung finden sich hier: http://www.librarytechtonics.info/archives/2005/10/paul_ngyuen_scr.html.

Seite als JPEG, DjVu-Java-Applet, als GIF, PDF sowie OCR-Datei aufzubereiten, ist Open Source und auf der Sourceforge-Seite als scribesw-Projekt[20] erhältlich.

Zur Vorstellung des Projektes rief Kahle den im *Golden Gate Club* nahe seines *Archive* versammelten Gästen vollmundig zu: „Lasst uns dem Volk seine Bücher zurückgeben." Im Blog zur OCA-Premiere beschrieb Kahle sein Ziel etwas ausführlicher. „Ist Open Content der nächste Schritt in der Tradition von Open Source und Open Network?" fragte der Computerwissenschaftler. „Viele Menschen scheinen das zu denken (und wäre es nicht wunderbar?). Wir arbeiten mit Büchereien, Regierungseinrichtungen, Archiven, Technologie und Webfirmen gemeinsam am gleichen Ziel: Es ist an der Zeit, mehr großartiges Material im Internet zu haben, das offen und gratis zugänglich ist." (Kahle 2005)

8 Suchmaschinen verschmelzen mit Online-Archiven – und werfen neue Probleme auf

Womit man wieder bei der Frage nach den Urheberrechten anlangt. Beim Copyright prallen die Visionen für im wahrsten Sinne des Wortes „offene Quellen" des Brewster Kahle auf die kommerziellen Interessen großer Internetfirmen wie Yahoo!, Google, Microsoft und Amazon sowie die Belange von Verlagen und Autoren. Sie alle haben in jüngster Vergangenheit ihre eigenen Konzepte vorgestellt, wie einem Internetnutzer Zugang zu digitalem oder nachträglich digitalisierten Inhalten gewährt werden soll. Und zwar in erster Linie, um Produkte zu bewerben und zu verkaufen.

Den größten Wirbel verursachte Google, als das Unternehmen im Dezember 2004 ankündigte, die Sammlungen der Büchereien an den Universitäten Harvard, Stanford, Michigan, Oxford und der *New York Public Library* zu scannen. Das so genannte „Google Print Library Project" soll Nutzern der Suchmaschine die Suche in Büchern nach Stichwörtern erlauben, als ob sie nach einer Webseite suchten, und dann relevante Passagen online anzeigen. Stanford alleine verfügt über 8,5 Millionen Bände, Michigan über rund 7 Millionen, die in den kommenden sechs Jahren erfasst werden sollen. Für Google ist das eine Verschlagwortung mit der Technik des 21. Jahrhunderts, die von der „Fair Use"-Klausel des amerikanischen Copyrights gedeckt ist. Danach ist es erlaubt, in Auszügen zu zitieren, ohne sich die vorherige Zustimmung der Verlage oder Autoren einzuholen. Kein Internetnutzer wird den vollständigen Text eines gescannten Buches bei Google aufrufen können.

Die Verlagswelt sieht das entscheidend anders. Erst warnten Interessenverbände die kalifornische Firma in geharnischten Diskussionen und Briefen, dann reichte die Autorengewerkschaft *Authors Guild* im September 2005 eine Sammelklage ein, gefolgt im Oktober von einer Klage der *Association of American Publishers (AAP)*, der mehr als 300 US-Verlage angehören. Im Kern lehnen die Verlage und Autoren Googles Argumentation des Scannens und der auszugsweisen Wiedergabe als „Fair Use" ab,

20 Siehe http://sourceforge.net/projects/scribesw/.

da die Urheberrechtsinhaber nicht nach ihrer Zustimmung gefragt werden, bevor die Suchmaschine ihre Werke digitalisiert und auf ihren Servern speichert. Schlimmer noch, so die Kritiker: Google verkaufe Anzeigen neben den Suchergebnissen an Inserenten und verdiene so schlussendlich am geistigen Eigentum anderer Geld. „Autoren und Verleger wissen, wie nützlich Google als Suchmaschine ist und dass die *Print Library* eine hervorragende Ressource sein könnte. Aber Tatsache bleibt, dass Google mit seinem gegenwärtigen Plan Millionen von Dollar aus dem Talent und geistigen Eigentum von Autoren und Verlegern schlagen will", erklärte AAP-Präsidentin Patricia Schroeder zur Begründung.

Ironischerweise ist der zweite Teil von Googles Archivierungs-Plan namens „Google Print Publisher Program" den Verlagen willkommen. An diesem bereits im Oktober 2004 gestarteten Programm, das inzwischen in „Google Book Search" umbenannt wurde, nehmen fast alle namhaften Verlage in den USA und Großbritannien teil. Sie können bestimmen, ob sie einen Titel in elektronischer Form an Google übermitteln und welche Passagen bzw. Seiten ein Google-Nutzer zu sehen bekommt (aber nicht ausdrucken kann), wenn er nach Stichwörtern sucht. Hat er Interesse, kann er das Buch mit ein paar Klicks kaufen, und Google verdient an den Kontext-Anzeigen Geld – die Grundlage seines Geschäftsmodells. Im Herbst 2005 schaltete Google diese Kategorie in acht europäischen Seiten live. Ähnlich verhält es sich bei Amazons „Search Inside the Book"-Angebot, das inzwischen auch auf dessen deutscher Seite erhältlich ist: Wer nach einem Stichwort sucht, findet im Katalog des Unternehmens oder in dessen Suchmaschine A9.com sogar Fußnoten in lieferbaren Büchern.

9 Die Debatte um Opt-In oder Opt-Out – ein Versuch der Landnahme?

Googles weitaus ehrgeizigeres Bücherei-Projekt weist einen fundamentalen Unterschied zu diesen Buch-Werbevehikeln und Brewster Kahles Projekt auf. Die *OCA* will nach dem Opt-In-Prinzip verfahren, wenn es sich um urheberrechtlich geschütztes Material handelt, während Google vom Opt-Out-Prinzip ausgeht. Das Unternehmen setzte sein Scan-Projekt für drei Monate aus, um Verlegern die Gelegenheit zu geben, Listen von Büchern zu übermitteln, die sie ausgenommen sehen wollen, und nahm die Digitalisierung der Stanford-Sammlung im November 2005 wieder auf.

Ohne *opt-out*, argumentieren Google-Manager, ließe sich die Katalogisierung von Büchern im zweistelligen Millionenbereich nicht bewältigen. „In Zukunft wird nur das gelesen werden, das online ist", sagte Jim Gerber (2005), Googles Direktor für Content-Partnerschaften, dem *Economist*. „Was nicht online ist, existiert nicht." Google–CEO Eric Schmidt (2005) präzisierte in einem Gastkommentar im *Wall Street Journal* kurz nach Eingang der Authors-Guild-Klage, wieso Autoren und Verlage für die Katalogisierung von Titeln, deren Copyright noch nicht abgelaufen ist, dankbar sein müssten:

> „Nach gängigen Schätzungen sind weniger als 20 Prozent aller Bücher noch lieferbar, und nur ungefähr 20 Prozent sind [...] in der Public Domain. Damit bleiben erstaunliche 60 Prozent aller Bücher übrig, die Verlage unserem Programm wider Erwarten zufügen könnten und die Leser nicht ausfindig machen können. Nur wenn wir sie scannen und Wort für Wort verschlagworten [...], können wir all diese verlorenen Titel vor dem Vergessen retten, und zwar so umfassend, dass Google Print eine Ressource wird, die die Welt verändert [...]. Man stelle sich nur einmal die kulturellen Effekte vor, wenn man zig Millionen bisher nicht zugängliche Bände in einen gewaltigen Index stellt, so dass jedes Wort, von Jedermann, ob arm oder reich, ob Stadt- oder Landbewohner, ob in der ersten oder dritten Welt, in jeder Sprache suchbar ist – und obendrein noch kostenlos." (Schmidt 2005)

Mit anderen Worten: Die positiven kulturellen und wirtschaftlichen Effekte eines Internetkatalogs all jener Werke, auf die Kahles Archiv aus Copyright-Gründen verzichten muss, würden die möglichen Gefahren von Urheberrechtsverletzung oder Diebstahl weit überwiegen. Hier nehmen auch der OCA-Partner und Verleger Tim O'Reilly sowie Juraprofessor Lawrence Lessig Google in Schutz. „In Vergessenheit zu geraten, ist die größte Gefahr für einen Autor", schrieb O'Reilly (2005) in der New York Times, und Lessig (2005) verurteilte die Klagen als „dreisteste Landnahme in der Geschichte des Internets, die Innovationen weitreichend zu behindern droht."

10 Kommerzielle Anbieter bauen konkurrierende Archivmethoden auf

Während die Parteien auf ihre Gerichtsverhandlung warten, oder einen Vergleich aushandeln, haben andere Marktteilnehmer ihre eigenen Versionen von Open Content gestartet. Microsoft unterstützt Kahles *OCA* nicht nur mit fünf Millionen Dollar, sondern will die gescannten Bücher zugleich seiner Suchmaschine *MSN Search* einverleiben. Yahoo! verspricht sich von der OCA-Zusammenarbeit ebenfalls einen gewissen Heiligenschein-Effekt als Verteidiger des freien Zugangs, und sieht es als Teil seiner Vision namens FUSE (für *Find, Use, Share and Expand*), die Wissen möglichst umfassend katalogisieren will. Für kommerziell ausgerichtete Portale sind Bücher und wissenschaftliche Literatur ebenso wichtige Magneten, um Nutzer auf ihre Seiten zu locken, wie Blogs, Fotodateien oder Podcasts.[21]

Verlage wiederum experimentieren mit ihren hausgemachten Strategien, um ihre Werke hinter elektronischem Schloss und Riegel zu behalten. Bertelsmann-Tochter Random House will Bücher seitenweise verkaufen (Applebaum 2005), und der

21 Eine Zusammenfassung von Yahoo!s Ideen finden sich in John Battelles (2005) *Searchblog*, in dem er eine Diskussion mit Yahoo!-Manager Jeff Weiner wiedergibt.

Börsenverein des Deutschen Buchhandels stellte auf der Frankfurter Buchmesse im Oktober 2005 sein eigenes Konzept der Web-basierten Volltextsuche vor. Dabei wären die Titel als PDF-Dateien von jedem Verlag auf einem eigenen Server gespeichert oder bei einem externen Dienstleister gehostet. Um die 3 000 Euro pro Server pro Jahr und um die 10 bis 100 Euro pro Buch würde das einen Verlag kosten, schätzt einer der Väter der Idee, der Stuttgarter Verleger Matthias Ulmer. Hat ein Verlag seine eigenen Werke erst einmal digital aufbereitet, ließen sich Rahmenverträge mit Amazon oder Google aushandeln, damit die Suchmaschinen fündig werden.[22]

11 Fazit: Open Content lässt künftige Generationen „auf den Schultern von Riesen stehen"

Eines droht auf der Strecke zu bleiben bei diesem noch lange nicht ausgefochtenen Machtkampf zwischen Vertretern der alten und neuen Wirtschaft, bei dem zudem europäische Ressentiments gegen die Dominanz amerikanischer Konzerne mitschwingen.[23] Wenn man Ideen, Fragmente oder Geistesblitze nicht ausfindig machen oder für eigene Schöpfungen verwenden kann, obwohl die heutige Technik es ermöglicht, aber überkommene Gesetze oder monetäre Verteilungskämpfe es verhindern, enthält man letzten Ende der nächsten Generation geistige Nahrung vor. Neben dem streng umzäunten Garten des Copyright muss es eine Spielwiese anderer Modelle geben, die sich am mittelalterlichen Modell der Allmende orientiert.

Was Brewster Kahles *Internet Archive*, die *Wayback Machine* und jetzt die *Open Content Alliance* im Blick haben, ist eine solche weitgehend offene digitale Kultur, wie sie die so genannte Wissens- oder Informationsgesellschaft verdient hat. Im akademischen Bereich gehen Projekte wie die *Public Library of Science* mit gutem Beispiel voran.[24]

Wer in den kommenden Monaten und Jahren auf den wachsenden Bestand an offen verfügbaren Ressourcen der *OCA* zugreift, kann Titel und Dateien aus der Public Domain weiter verwerten und daraus neue Werke schaffen – und sie sogar kommerziell verkaufen. Open Content lässt sich dann genauso „verkosten" oder „sampeln" und neu kombinieren wie heutige *Web Services*, die etwa *Google Maps* mit Wohnungsinseraten oder neuesten Sonderangeboten um die Ecke verknüpfen. „Wenn jemand [von euch] aus den Stücken der Public-Domain-Sammlung ein Buch drucken

22 Umfassende Dokumente zum Konzept und seiner Realisierung sowie möglichen Kosten finden sich auf der Webseite des Börsenvereins unter: http://www.boersenverein.de/de/69181?rubrik=86662

23 Siehe etwa die von Frankreich angeregten Pläne, ein eigenständiges europäisches digitales Archiv von Manuskripten, Büchern, Fotos und Noten anzulegen. Sechs Mitgliedsstaaten der EU haben bereits angekündigt, die Bestände ihrer Nationalbibliotheken digital zu erfassen: http://www.dwworld.de/dw/article/0,1564,1566717,00.html.

24 Das University-of-California-System etwa zahlt für den traditionellen Online-Zugang zu Fachzeitschriften und Datenbanken nach Auskunft des Leiters seines digitalen Bücherei-Programms, Daniel Greenstein, rund 24 Millionen Dollar Lizenzgebühren im Jahr. Details unter: http://plos.org/.

und binden will und es dann auf Amazon.com anbietet", träumt Brewster Kahle von seiner weltumspannenden Bücherei, „habt Spaß! Wenn man es als Audiobuch aufnehmen und ins Web stellen will – nur zu. Wir werden es sogar hosten. Auf den Klassikern der Menschheit aufzubauen, soll eine Riesenparty werden."

Literatur

Applebaum, S. (2005), 'Random House, Inc. Announces Business Model for Online Viewing of Books'. http://www.randomhouse.com/trade/publicity/pdfs/OnlineViewingRH1.pdf [29. Jan 2006].

Battelle, J. (2005), 'Missions and Visions'. http://battellemedia.com/archives/001473.php [29. Jan 2006].

Gerber, J. (2005), 'Pulp friction', *The Economist* **12. Nov 2005**, S. 63. http://www.economist.com/business/displaystory.cfm?story_id=5149499 [11. Feb 2006].

Kahle, B. (2005), 'Announcing the Open Content Alliance'. http://www.ysearchblog.com/archives/000192.html [29. Jan 2006].

Kesmodel, D. (2005), 'Lawyers Delight: Old Web Material Doesn't Disappear', *Wall Street Journal* **27. Jul 2005**, S. A1. http://www.prnewsonline.com/legalpr/case_wayback.html [11. Feb 2006].

Lessig, L. (2005), 'Google's Tough Call', *Wired* **13**(11). http://www.wired.com/wired/archive/13.11/posts.html?pg=8 [29. Jan 2006].

Lyman, P. und Varian, H. R. (2003), How Much Information?, Studie, University of California, Berkeley. http://www.sims.berkeley.edu/research/projects/how-much-info-2003/ [29. Jan 2006].

Mills, E. (2005), 'Google ETA? 300 years to index the world's info', *CNET News.com* . http://news.com.com/Google+ETA+300+years+to+index+the+worlds+info/2100-1024_3-5891779.html [29. Jan 2006].

O'Reilly, T. (2005), 'Search and Rescue', *New York Times* **28. Sep 2005**, S. A27. http://radar.oreilly.com/archives/2005/09/ny_times_op_ed_on_authors_guil.html [11. Feb 2006].

Prelinger, R. (2001), 'An Informal History of Prelinger Archives'. http://www.panix.com/~footage/shorthistory1.html [29. Jan 2006].

Schmidt, E. (2005), 'Books of Revelation', *Wall Street Journal* **18. Okt 2005**, S. A18. http://googleblog.blogspot.com/2005/10/point-of-google-print.html [11. Feb 2006].

Squeo, A. M. (2005), 'Oh, Has Uncle Sam Got Mail', *Wall Street Journal* **29. Dez 2005**, S. B1. http://online.wsj.com/public/article/SB113581938626033499-xNP7F7iqAatGMjivCNMuy6GOH2M_20061229.html?mod=blogs [11. Feb 2006].