世界のウェブページを時間軸に沿い集め、公開する
ウェイバックマシーンとは何か

立命館大学先端総合学術研究科

山口 和紀

要旨：
世界中のウェブ上の記録は、いままさに消失しつつある。この消失に対して、いかに対処すべきかは十分な議論がない。そこで本稿では、もっとも先駆的なウェブアーカイブの一つであるインターネットアーカイブのウェイバックマシーンに着目し、その歴史・経営・技術・法律などを多角的に検討した。ウェイバックマシーンはウェブ上の情報を収集し公開することによって、ウェブ上の情報の散逸を防いでおり、その試みは一定の成功を収めている。とくに許諾なしにすべてのウェブサイトのアーカイブを試みている点は、世界的にも稀であり、貴重である。ただし、技術的・法的な課題や、経営的な課題もある。ウェイバックマシーンをどのように捉え、日本社会の中に位置付けるのかは今後の課題である。

キーワード：
アーカイブ、ウェブアーカイブ、インターネットアーカイブ、ウェイバックマシーン

1. はじめに
1-1. 「記録が残されていない時代」に
インターネット上の記録は、次々と書き換えられ、失われていく。その速度はときに「紙」やアナログ媒体に保存された記録よりも早い。いまや、多くの情報がインターネット／ウェブ上にのみ記録され、印刷物やアナログ媒体に記録されることのほうが少ない（時実［2016:490］）。それを残すための万全な手立ては講じられてこなかった。

この状況について「このままでは 21 世紀は『記録が残されていない時代』になってしまう恐れがある」（時実［2016:490］）とされる。

中井［2022］は、2000 年代に盛んになったブログサービスが 2010 年代から 2020 年代にかけて急速に終了していったことを指摘している。そこには病や障害を持った当事者の記録が残されていたが、それらの多くは消えてしまった／消えていくと考えられる（中井［2022:60］）。

また、日本の国立国会図書館の調査では、国の機関のサイトのウェブページ（正確には URL）の 5 年残存率は「40%」であった（国立国会図書館［2016］）。ある国の機関が発行したウェブページの URL があったとして、5 年後にその URL が消えてしまいアクセスできなくなる割合は、過去のデータでは 6 割に上ったということである。

当事者の記録も、国の機関の記録も放っておいては残らないのである――国のウェブサイト

上の記録については現在、国立国会図書館がアーカイブしている。大量かつ重要な情報が、いままさに消え去っていっている。このことをどのように考え、対処しようとするのか／しうるのか。これは単にその保存を行おうとする機関や組織だけの問題ではない。この社会がそれをどのように捉えるのかが問われる。本稿はそれを捉える試みの端緒を開くためのものである。

そのために、これまで行われてきたことを振り返ることにより、不足する部分／対処すべき部分を切り分ける。

1-2. ウェブアーカイブとはなにか

本稿が議論の対象とするのは「ウェブアーカイブ★01」と呼ばれるものである。まずウェブアーカイブについて本稿における定義を明確にしておきたい。

そもそもアーカイビングは、記録された媒体の性質によって必然的にその収集・公開方法が異なっている。

これまで盛んに行われてきたのは「紙」媒体に記録された資料のアーカイブである。これは「紙」の劣化を防いだうえで、それを整理し、公開に供するものである。他のアナログ媒体でも同様である。

そのほかに「紙」などのアナログ媒体に記録された資料を「デジタル」化することによって、アーカイブしようとする試みもある。デジタルアーカイブなどと呼ばれるこの分野は、近年進展が著しい。デジタルデータは基本的に劣化しないため、永続的な保存と提供が原理的には可能という大きなメリットがある。

本稿が述べようとするのは、それらとは異なるものである。それは作成当初からデジタルで記録され流通することを前提としている資料等（Born digital★02）のアーカイブである。直接的に言えば、「ブログ」や「ホームページ」の書き込みのアーカイビングである。これはウェブアーカイブと呼ばれている。

アナログをアナログのままアーカイブすること、アナログ媒体をデジタル化すること、最初からデジタルであるもの――すなわちモノとして存在していないもの――をいかにアーカイブするかはそれぞれ基本的な部分で方法や位置付けが異なっている。

1-3. 本稿の位置付け

ウェブアーカイブについての研究は、まず技術に関わるものがある。それはどのようにインターネット上の情報を集め、保存し、公開するのかといった技術的な側面についてである。そうした技術的な議論や発展の経緯は、前田ら［2017］や原田［2008］に詳しい。

次に、法律に関するものがある。単に技術があるだけでは、ウェブアーカイブは成立し得ない。ウェブアーカイブの形成期の論点については新保［2008］、そのうちの著作権に関する議論は東［1999］に詳述されている。また、国や機関によってもウェブアーカイブを支える法枠組みは異なるが、そうした差異については日米を比較した山口［2022］や、日本のウェブアーカイブの形成期における法的な議論に焦点を当てた山口［2023］に詳しい。

その他に、ウェブアーカイブを支える組織や一つひとつのウェブアーカイブの実践に焦点を当てた研究がある。例えば、世界的な協働を目指す機関としてInternational Internet Preservation Consortium（以後、IIPC★03）があるが、それを中心として何が行われているのかについては柊ら［2008］の研究がある。

本稿は、ウェブアーカイブの代表的な事例を取り上げ、その事例がどのようにウェブアーカイビングを行っているのかを検討する。技術、法律、運営のいずれの側面も検討に含め、多角的に検討する。

検討の対象として取り上げるのは「Wayback Machine（以後、ウェイバックマシーン）」である。これは世界で最も有名なウェブアーカイブである。

なぜウェイバックマシーンを取り上げるのかというと、それが代表的なウェブアーカイブであるだけでなく、ウェブアーカイブの黎明期に立ち上がった経緯から各国のアーカイブのモデルとなってきたからである。

例えば、後述するようにIIPC加盟機関の多くが使っているインターネット資料の収集ソフトは、ウェイバックマシーンの運営元が開発の中心的役割を果たし、作成され保守され続けてきた。

1-4. ウェイバックマシーンの概要

ウェイバックマシーンは、アメリカの非営利組織 Internet　Archive（以後、インターネットアーカイブ）が運営するウェブアーカイブである。その設立の経緯と展開については後述する。

様々なウェブサイト、コンテンツが収集されており、誰でも閲覧が可能な状態になっている。後述するように機械的に収集がなされており、もとの状態のまま閲覧することが出来るのが特徴である。

ただし、利用規約（Internet Archive［2014］）には「アーカイブのコレクションへのアクセスは無償で提供され、学問や研究の目的にのみ利用可能である＝筆者訳（原文：Access to the Archive's Collections is provided at no cost to you and is granted for scholarship and research purposes only）」と述べられている。

このウェイバックマシーンは日本国内からも利用が可能なサービスである。サイトアドレスは、「archive.org/web」であり、「ウェイバックマシーン」などと検索すれば上位に表示され、アクセスが可能である。

ウェイバックマシーンは、世界中のウェブサイトをアーカイブしている。ウェイバックマシーンのトップページには、収集されたウェブページの数が表示されている。執筆時点では「more than 808 billion（8080 億以上）」と書かれている。

例えば、生存学研究所ＨＰとして利用されている http://www.arsvi.com/ を指定すると、2002 年 5 月 25 日から執筆時点（2023 年 6 月）までで 352 回の保存が行われていると表示される。

そのうち、もっともはじめの保存データを見てみると、次のように、その保存された瞬間の HP の様子が再現されている。ウェイバックマシーンは、あるウェブページに対して、時系列的な保存を行っているのである。

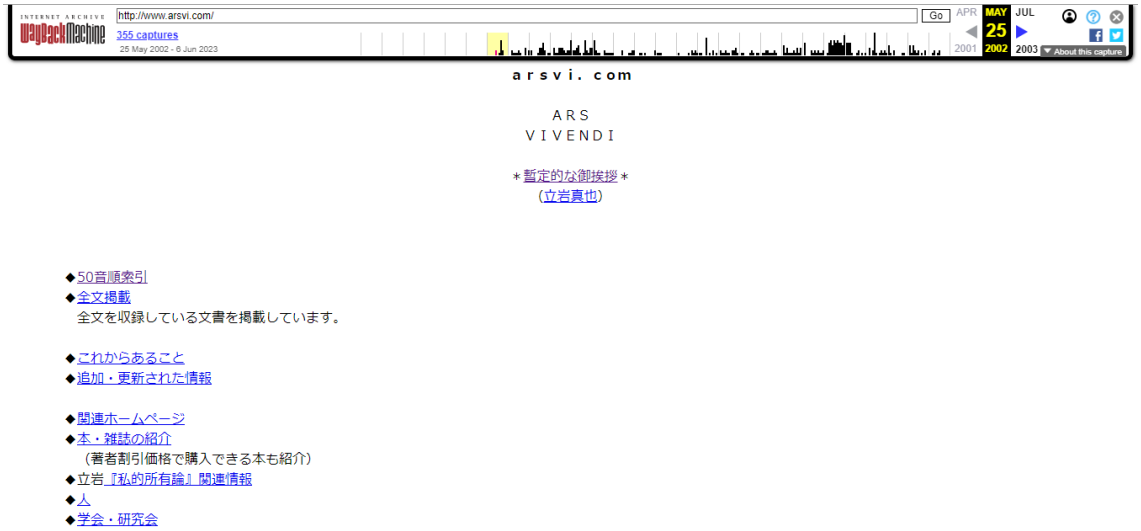

図 1. ウェイバックマシーンに保存された 2002 年 5 月 25 日時点の arsvi.com のトップページ

2. ウェイバックマシーンとは何か

2-1. いかに始まり、続いてきたのか

まず、ウェイバックマシーンの試みがいつどのように始まったのかを見る。ウェイバックマシーンの歴史を説明するために、その運営主体であるインターネットアーカイブの説明から始める。

インターネットアーカイブは、1996 年にブリュースター・ケール（Brewster Lurton Kahle, 1960 年 10 月 22 日〜）によって設立された。

ケールはマサチューセッツ工科大学（以後、MIT）を 1982 年に卒業した。学士号は計算機科学・工学であった。

ケールは、インタビューで次のことを語っている。

——インターネット・アーカイブのような「図書館」を作ろうと最初に思いついたのはいつでしょう。

BK：20 歳のときだから、1980 年のことだ。僕はすでに大学生になっていた。問題は、コンピュータをどう扱うかということだった。1960 年代的な新しい考え方が当時はまだ残っていた。そして僕がいちばん使い方を分かっていたのはコンピュータだった。当時は二つのことしか考えられなかった。一つは、人々のプライバシーを守ること、ようするに盗聴されることなく人々が電話で話し合えるようにすること。もう一つは巨大な図書館、古代アレクサンドリア図書館のバージョン 2.0 を作ること。でも、後者は分かりやすすぎるテーマだから、きっと他の人がやってくれるだろうと思っていたんだ。

そこでまずプライバシーのほうに取りかかることにして、コンピュータチップをプライ

> バシーのためだけに使う方法を学んだ。でもそれは、僕が助けたいと思う人たちの助けにはならないことに気付いた。というのも、当時は安くコンピュータチップを作れなかったから、そういうチップを作っても大企業や軍やマフィアを助けるだけだった。僕はどれも助けたくはなかったからその仕事は止め、「図書館」を作ろうということになった。あれから 30 年になるけど、僕は今でもその「図書館」を作ろうとしているんだ。（ブリュースター・ケール［i2011］）

ケールは MIT 在学中の 1980 年にこの「図書館」を構想したというが、卒業後まず「Thinkng Machines（シンキングマシンズ）という会社の社員になった。この会社は、MIT で開発されていた「コネクションマシン」の商用化を目指していた。ここでケールは、WAIS（Wide Area Information Server）を開発し、「WAIS 社」を立ち上げ、1995 年に売却する（仲俣［2011］）。その売却額は 1,500 万ドルであった（Hardy［2009］）。

その後 1996 年、ブルース・ギリアット（Bruce Gilliat，1959 年 5 月 30 日〜）と共に非営利組織としてインターネットアーカイブを創設する。

このインターネットアーカイブが今日までウェイバックマシーンの運営元となってきた。

同時にケールは、営利企業として、「Alexa Interet（アレクサインターネット）」も設立している。このアレクサインターネットは 1999 年に Amazon.com に売却され傘下となるが、その売却額は 2 億 5,000 万ドルであった（Hardy［2009］）。

このアレクサインターネットは、ウェブページに対するアクセスを解析するツールを開発していた会社である。ウェイバックマシーンとアレクサインターネットの関係については後述する。

アレクサインターネットが収集したウェブサイトのデータは、インターネットアーカイブがアーカイブコンテンツ（アーカイブされた資料）として用いていた（Mohr［2014］）。

このようにウェイバックマシーンが始まったのは 1996 年であるが、同時期にウェブアーカイブを開始した組織としては「オーストラリア国立図書館」、「スウェーデン国立図書館」があった（国立国会図書館［2020］）。ウェイバックマシ―ンは、最も先駆的なウェブアーカイブの組織の一つであった。

ただし、ウェイバックマシーンが現在のように一般公開がなされるようになったのは 2001 年のことである。その時点で 100 テラバイト程度の保存されたデータがあり、100 億ページ程度アーカイブされていた（Internet Watch［2001］）。なお、近年の動きについては時実［2016］などに詳しい。

2-2. 資金をいかに集め、使うか

独立した非営利組織としてのインターネットアーカイブはいかに資金を集め、どのように使っているのだろうか。

ブリュースター・ケールは次のように語っている。

――インターネット・アーカイブの運営資金についておしえてください。

BK：最初のお金は自分が出した。AOL のスティーブ・ケースやアマゾンのジェフ・ベゾスが前の会社を買収してくれたおかげで、僕は大金を手にし、インターネット・アーカイブを始めることができたんだ。でもそのうち、他の図書館と連携できることに気がついた。僕らは、すべてのパブリッシャーをオンライン化させたのと同じやり方で、図書館と連携することを望んでいる。すべての図書館を前進させたいんだ。[…]

インターネット・アーカイブの資金は、世界中の国立図書館および大規模な大学や公共図書館から来ている。ワールド・ワイド・ウェブのアーカイブ化（Wayback Machine）か、本のデジタル化のいずれかのためだ。一部の資金は、ベンチャーキャピタルみたいな財団からも得ている。ベンチャーキャピタルは、こちらがお金を返すことを望んでいない。ただ「良いこと」をやってほしいということで、僕らのようなプロジェクトの立ち上げを支援しているんだ。（ブリュースター・ケール［i2011］）

その具体的な内容について、2019 年の年次財務報告書（Internet Archive［2019］）をもとに、インターネットアーカイブの財務状況を概観する。Internet Archive［2019］は、Form990 と呼ばれる資料で、アメリカの非営利組織が提出する財務報告書である。ただし、インターネットアーカイブは、本稿が扱うようなウェブページのアーカイビングだけではなく、映像資料や図書資料のアーカイビングも行っている。ここで示すのはその全体の財務状況であって、ウェブアーカイブ単体の収支ではない。

当該年度の総収入は、3,671 万米ドルである。総経費は、3,725 万米ドル。 1 ドルを 130 円とすると、総収入は 47 億円、総経費は 48 億円程度になる。なお比較対象として、日本の国立国会図書館は令和 5 年度の当初予算で約 198 億円を計上している（国立国会図書館［不詳］）。その役割がまったく異なるため比較の対象として適切ではないものの、予算規模としては国立国会図書館の 4分の 1 程度であるということになる。

その収入の約 80%（29,117,560 米ドル）を占めるのが、寄付・助成（Contributions）である。その次に、事業収入（Program Service）が 20%（7,563,418 米ドル）程度を占めている。ただし、この割合については年度によって多少の変動がある。2019 年はやや寄付・助成による収入が大きい。

さらに細かく寄付・助成の内訳をみると、まず寄付では政府機関からの助成（Government grants(contributions)）による収入が、146,366 米ドル計上されている（Internet Archive［2019:9］）。次に、それ以外の寄付・助成金として 28,971,194 米ドルあり、現金以外の寄付・助成金として 221,612 米ドルとある（Internet Archive［2019:9］）。

このようにその収入のほとんどが助成・寄付によって賄われている。その助成の内容について

いくつかを紹介する。2017 年、インターネットアーカイブは博物館・図書館サービス機構（Institute of Museum and Library Services, IMLS）からの助成（Bush 21st Century Librarian grant）を受け、アメリカ全土の公共図書館に対して地域情報のウェブアーカイビングの専門知識のトレーニング、サポートを行っている（Internet Archive［2017］）。この助成には、2020 年にも追加の助成がなされている。

インターネットアーカイブの HP 上にある「About the Internet Archive」という記事（Internet Archive［不詳］）によれば、全米人文科学基金（The National Endowment for the Humanities, NEH）やアメリカ国立科学財団（National Science Foundation）などからも助成を受けている。

次に事業収入の内訳をみると、アーカイブ料金（Arcive Fees）として 551,261 米ドル、書籍のスキャン（Book Scanning）として 2,660,481 米ドル、サービス料金（Fees for Service）として 4,351,676 米ドルとある（Internet Archive［2019:9］）。

最後に、経費（Functional Expenses）は多岐に渡るのだが、とくに大きな経費は 3 つである（Internet Archive［2019:10］）。第一に、給料等である。第二に、国内の組織や政府機関に対する助成金事業である。第三に、雇用に関わらないサービスへの経費である。

スタッフの数については 2011 年の時点で「プログラマ、管理者、ライブラリアンをあわせて 50 人。それに 6 カ国、23 の都市にあるスキャニング・センターで働く人が、それとは別に 150 人いる」（ブリュースター・ケール［i2011］）とされている。

このように見ると、基本的にはその収入のほとんどは寄付や政府の助成金であり、あくまでも政府の機関ではなく独立した非営利組織として運営がなされていることがわかる。各国のウェブアーカイブが大学や国立図書館によるものである場合が多いことを考えれば、特殊な事例である。

2-3. どのような仕組みで集め、保存し、公開するのか

ウェイバックマシーンの仕組みについては、Mohr ら［2004］や前田・大山［2017］に詳しい。前田・大山［2017］は「ウェブアーカイブを支える技術」と題された論文である。

仕組みには 3 つの側面がある。第一に、資料を集める仕組みである。第二に、保存する仕組みである。第三に、それを利用者が閲覧できるようにする仕組みである★04。

第一の仕組みは「Crawler（以後、クローラ）」と呼ばれる。Crawl は「這い回る（はいまわる）」を意味する動詞であり、インターネットを這い回って情報を集めていくシステムのことを意味する。

このクローラはウェブアーカイブだけでなく、Google などの検索エンジンにも利用されている技術である。ウェイバックマシーンで主に用いられるのは「Heritrix」と呼ばれるクローラである（前田・大山［2017］）。ただし、ウェイバックマシーンは Heritrix によるクローラのみに情報の収集を頼っているわけではない。補足すれば、ウェイバックマシーンは単一のシステムを指す言葉ではなく、いくつかのシステムの複合としてのサービスを指していると言いうる。

そもそも、ウェイバックマシーンは 1996 年頃からの収集データを保持しているが、その時期

のデータはインターネットアーカイブ自身が収集したものではない。

アレクサインターネットという 1996 年に前述のブリュースター・ケールとブルース・ギリアットによって設立されたウェブページのアクセスに関する分析ツール作成会社が収集したものである（Mohr ら［2004］）。

当初、インターネットアーカイブはアレクサインターネットから収集データを譲りうける形でアーカイブを形成していたのである。

2003 年には、インターネットアーカイブ自身が収集を行うべくオープンソース方式によるクローラの開発に着手した（前田・大山［2017］）。それが Heritrix である。Heritrix は「（女性の）遺産継承者、相続人」を意味する。オープンソースによる開発としたのは、ウェブアーカイブに関心を持つ機関との協働を可能にするためであった（Mohr ら［2004］）。

クローラはまず起点となるページにアクセスし、そのページを複製したうえでリンクを解析し、リンク先にも同様の処理をしながら、指定した範囲全体のコンテンツを集めきるまで処理を繰り返す（前田・大山［2017:75］）。

前田・大山［2017:75-74］によれば、IIPC（★03 を参照）の加盟機関の多くがクローラに Heritrix を採用している★05。前田・大山［2017:75:74］の表に記載された 45 機関のうち、33 機関が Heritrix を採用している。

ただし、後述するように全世界のウェブ上のデータすべてを常に保存することはできない。あくまでも、定期的なクローリングによってデータを収集しているに過ぎない。

第二に、保存する仕組みがある。これは専門用語では「保存ファイルフォーマット」と呼ばれる。

ファイルには保存の形式がある。例えば、テキストファイルであれば「txt」拡張子によって保存されるし、表計算ソフトであれば「csv（comma-separated values）」形式によって保存される。

同様にウェブアーカイブに適したファイルの保存形式が考案され、用いられている。広く用いられているのが「WARC」と呼ばれるファイルフォーマットである。WARC は、もともとインターネットアーカイブが開発した「ARC」フォーマットを基礎として、長期保存やデータ交換により適したものに改良した形式であり、2004 年に IIPC によって拡張設計がなされた（前田・大山［2017:75:75］）。

WARC にはコンテンツそのものの情報も記録されるが、そのコンテンツをいつ、どのように収集したのかという情報もセットで記録されている（前田・大山［2017:75-76］）ことが特徴と言いうる。

第三に、利用者の閲覧の仕組みである。WARC フォーマットで保存されたコンテンツを閲覧するためのソフトウェアが存在する。その一つが「Wayback」である。ウェイバックは、インターネットアーカイブの「ウェイバックマシーン」をベースとして作成されたオープンソースソフトウェアである。2013 年には開発の主体が IIPC に移管され、2014 年 9 月に「OpenWayback」として公表されている（前田・大山［2017:75-77］）。

こうした技術によってウェイバックマシーンは支えられている。

2-4. 収集の範囲

ウェイバックマシーンは基本的にすべてのウェブページを集めようとしている。それは米国のウェブサイトでも、日本のウェブサイトでも関係なく収集するということである。実際にそれはなされている。

ただし、ウェイバックマシーンは、インターネット上の資料のすべてを網羅的に集めているわけではない。

そもそもすべての情報を集めることは、莫大な（計算・通信における）資源を必要とし、ほとんど／明らかに不可能である。いかに効率的に、重要な情報を集め、保存する——保存することにも当然コストはかかる——ことができるのかという経済性の観点は、モノが存在しないウェブアーカイブにおいても重要である。

ウェイバックマシーンはあくまでも、クローラによって、継時的／時系列的にウェブサイトにアクセスし、保存しているに過ぎない。また、ウェイバックマシーンのクロールを拒否したり、公開データからの削除（除外）を申請することも可能である。そのため、当然ではあるが、残らないウェブページも多い。

管見の限り、ウェイバックマシーンが存在するウェブページをどの程度網羅的に収集しているかの推計を行った研究はなく、不明である。また、ある時点 A における状態と、それが変化した状態 B は別のコンテンツであるとすれば、それらを網羅的に収集する場合の保存すべき情報量は「爆発的増大」をすることになる。ゆえに、どの程度網羅的にアーカイブされているかを推計することは、かなり困難なことであると考えられる。

そこでここでは、どのようなウェブページが残りにくいのか／ウェイバックマシーンがクロールしないのかという論点についてのみ言及する。

そもそもクローラには、ウェブサイトへの巡回の順番やその頻度を決める機能が備わっている★06。Heritrix の場合は、フロンティア（Frontier）というコンポーネント（プログラムの要素）がその役割——巡回すべき URI をProcessing Chains（保存処理の仕組み）に渡す——を果たしている（前田ら［2017］）。

例えば、Heritrix は BdbFrontier と呼ばれるフロンティアを採用している（Alex［2018］）★07。BdbFrontier には「キュー」と呼ばれる巡回すべきウェブページのリストが保持されている（Alex［2018］）。そのキューには「予算（Budget）」が設定される。この予算は、そのキューがどの程度の「注意（Attention）」を受けるか、つまりどの程度の時間を使って巡回するかを規定する。予算がなくなればそのキューへの巡回は止まるが、予算が続く限りは巡回される。このようにしてクロールの順序、範囲は整理されている。

その「予算」——すなわちどの程度のコストをかけてクロールされるのか——は、そのキューの重要性によって定まる。「関心を引くキュー（つまりコストが低い）はより多くの注意を受け、関心を引かない（つまりコストが高い）キューはより少ない注意しか受けない＝筆者訳（原文：Thus, queues with more interesting (less costly) URIs get more attention, while those with less

interesting (more costly) URIs get less attention)」とされる（Alex［2018］）。

このような形でクローラにはウェブサイトに対する効率化／重み付けの機能が備わっており、重要なサイトはより多く巡回し収集するが、そうでなければ頻度は落ちることになる。場合によっては、そもそも巡回そのものがなされないウェブページも存在することになる。

また、そうした重みづけとは別に、技術的な制約からクローラが収集しにくいウェブサイトも存在する。まずは ID とパスワードが必要なウェブサイトである。ウェイバックマシーンはこうした制限がかかっているウェブページを基本的に巡回しない／しえない。ウェイバックマシーンも、基本的には一般のユーザと同じようにウェブページにアクセスしているに過ぎず、ID とパスワードを要求されるページにはアクセスのしようがない。

また別の側面では、固定の URL を持たず、ユーザの入力に対して随時に動的にデータベースからウェブサイトを生成するようなページも、データベースそのものをクローラは収集していないので、アーカイブされにくい。例えば、SNS などはアーカイブされにくい★05 という報告がある（Schafer ら［2021:132-133］）。すなわち、クローラ方式自体に限界が存在する――入出力の結果を拾い集めるよりも、入出力の関数やそれが参照するデータベースそのものを直接得た方がコストは下がる――のだが、最も効率よく／機械的にデータを収集する方法はクローラである。

すなわち、①限られた資源をいかに有効に活用するかという側面で優先度が低いウェブページがあり、②優先度が高くても技術的に収集しにくいウェブページもあるということになる。大きくはこうした要素がアーカイブのされやすさを規定している。

2-5. 「すべてを集める」ことを巡る法／倫理

単に技術が存在するだけでは、ウェブアーカイブは成立しない。

前項で述べた通り、ウェイバックマシーンのようなウェブアーカイブはクローラによってデータを収集する。それはいわば、公開された情報を外部から勝手に取りに行き、さらに公開している状態である。これが法的に許容されうるかが問題となる。

こうしたウェブアーカイビングと法律を巡る問題については、神保［2008］（「ウェブ・アーカイビングと法」）や山口［2022］・山口［2023］などに詳しい。とくに日本国内における議論は新保［2008］や山口［2023］に述べられている。

ウェイバックマシーンは、著作者の許諾なしに収集を行っている。これは米国著作権法におけるフェア・ユースに則るものである（塩崎［2019:4］，山口［2022］）。東によればフェア・ユースとは次のような概念である。

> 著作権法の目的とするところは、著作物のような「文化的所産の公正な利用」を図り、もって「文化の発展に寄与する」ことである（著作権法第 1 条）とし、著作物を知的財産（intellectual property）とみなして、財産権（property rights）を著作権者（copyright holder）

に独占的・排他的（exclusive）に付与している。この独占的な権利を認めて著作権者を保護することにより、創作意欲や経済的にも報われる可能性のインテンシブを著作権者に法的に約束することが、公平な社会という要請に応えることになるからである。

しかし、反面、この独占的排他権は絶対的な権利ではなく、社会・文化の発展に寄与するうえで、社会的公正の範囲で制限が加えられてしかるべきであると考えられる。つまり、「公正な使用（フェア・ユース）」という、ユーザーが著作者の承諾なしに著作物を使用できる範囲を法的に規定したのがフェア・ユースの法理（fair use doctorine）である。（東［1999:67］）

ウェイバックマシーンは、このフェア・ユースの概念を用いて、自らの収集を法律の枠組みの中に収めている。これは「オプトアウト方式」とも呼ばれる。オプトアウト方式とは、収集したコンテンツを基本的には公開し、削除の申立てがあった場合のみ公開を停止するという方式である。

塩崎［2019:4］によれば、こうしたフェア・ユースに基礎づけられた「オプトアウト方式」による運用が可能なのはアメリカやカナダであり、アメリカやカナダの大学ではウェブアーカイブを構築する事例が見られる。

日本の国立国会図書館が行っているウェブアーカイブの事業では、こうした「すべてのウェブページ」に対する著作者の許諾を得ない収集は行われていない。ウェイバックマシーンのように、全世界のウェブ上の資料を網羅的に集めようとする試み（バルク収集などと呼ばれる）は、世界中のウェブアーカイブのなかでも例外的なものであると言える。

3. いかに捉え、位置付けるか

3-1. ウェイバックマシーンをいかに捉えるか

ウェイバックマシーンは、消失しつつあるウェブ上の資料を、万全ではないとはいえ残し続けている。これをどのように見るべきだろうか。

現在の日本では民間の「ウェブサイト」をアーカイブすることは十分に／法的には行われていない。そのため、そうした情報のほとんどは消えて行ってしまうことになるが、ウェイバックマシーンはその一部分――「日本」のウェブサイトの収集の数は不明だが、膨大な数に上ることは明らかである――を現に集めているし、これからも集めていくだろう。

ウェイバックマシーンがあることによって、消失しなかった「日本」のウェブサイトは数多くある。その意義は大きく、日本の社会にとってもウェイバックマシーンは重要なものであるといいうる。ここまではまず重要である。

そこからどのように考えるかは、議論の余地がある。日本でもそうした「承諾を得ない」形での網羅的なアーカイブが模索されていたが、議論の過程においてその範囲を大きく限定したアーカイブ事業（国立国会図書館によるインターネット資料収集保存事業＝WARP）が立ち上がっ

たことは山口［2023］において述べた。すなわち、日本（の国立国会図書館による事業）では、ウェイバックマシーンのような網羅的収集は難しいという判断がなされたのである。

アメリカでそれが可能になったのは、本稿ですでに述べたように「フェア・ユース原則」の範囲に収まるという法的な解釈による部分が大きい。しかしそれだけではなく、そうした試みに対してアメリカ社会が経済的な援助を行ってきたこと、ブリュースター・ケールが会社の売却益を「全知識体系への全世界的アクセス」のために使ったこと――ケールが創設時の資金を調達したのは通販サイトとして知られる「Amazon」に会社を売却したことが大きい（ブリュースター・ケール［i2011］）――など、それを成立させた社会的土壌の側に注意を払うべきである。

それはウェブアーカイブも社会的な構造の中において成立しうるものであり、それがどのような役割を果たしうるのかは、社会関係の中で捉える必要があるということである。単に技術的／法的な側面だけに着目し、論を進めようとするのではなく、ウェブアーカイブという試みを社会全体がどのように捉え、動かそうとするのかを見る必要がある。この点については、別稿にて、さらに深く論じたい。

3-2. どのように位置付けるのか

ウェイバックマシーンが存在するとはいえ、日本国内のみならず、世界中のウェブサイトが残されないまま消失していっている。

記録を残すこと自体の有用性は誰も否定しない。しかし、ウェブ上のすべてを集め、公開することは一定の問題を含んでいる。例えば、その情報を発信した当人にとって残ってほしくない情報が残ってしまう、個人情報とそうでないものを見分けることは膨大なコストがかかると考えられることなどである。また、許諾を得ずに収集し公開する場合の著作権や補償の問題もある。手放しにすべての記録を残し、公開すべきだと主張をすることは難しいだろう。

ゆえにどこまでが可能かつ適切な範囲であるのかを見極め、社会的な合意を生み出す必要がある。日本におけるウェブアーカイブをいかに進展させるかという観点からは、ウェイバックマシーンはいかに捉えられるだろうか。

日本の国立国会図書館は、ウェブアーカイブ事業「WARP」の収集の範囲を徐々に広げてきた。例えば、本稿が述べるようなアーカイブとはやや異なるが、有料の電子書籍・電子雑誌のアーカイブ事業（有償等オンライン資料の制度収集の一部）も、2023 年より行われはじめた。

しかし、今日にいたるまで「民間」のウェブサイトのアーカイブは行われていない。どうしてそのような制度的な収集を公的機関のウェブサイトに留める方向に舵が切られたのかについては山口［2023］で分析を行ったが、その大きな要因は、国の機関が個々人の発信を集め、保存することは社会的な理解を得ることが難しく、著作権などの権利的な問題もクリアし難いという考えが主要なものであった。

ウェイバックマシーンは、アメリカで行われているウェブアーカイブである。ゆえに、日本のウェブサイトも収集しているのだが、あくまでも原則的にはアメリカ合衆国の法的な規制／基盤のもとにある。日本では困難であり、現時点では行われていない民間のウェブサイトに対する

許諾を得ない収集が、アメリカで「代行」されている状況である。

このような関係性の中で、国立国会図書館のウェブサイトのアーカイビング事業、国内で構想し設置しうる新たなウェブアーカイブ事業をどのように動かすのか／動かしうるのかを考えることは重要である。

第一に、ウェイバックマシーンがあるからそれで十分であると考えることができる。実際には収集は部分的なものにすぎないとはいえ、世界中のありとあらゆるウェブサイトを網羅的に収集するための要件をすでにウェイバックマシーンは満たしており、それを拡充させるように働きかけること／協働することがある。それによって不十分な点を補う、資金的な援助をするという方向性は考えられる。

第二に、ウェイバックマシーンでは全く足らないと考え、新たな事業／既存事業の拡充を目指すことも考えられる。ウェイバックマシーンはあらゆるウェブサイトの収集を行い、莫大な蓄積があるとはいえ、全世界のウェブサイトの完全に網羅的なアーカイブズを作成しているわけではない。それでは不十分であると考えるならば、新たに事業を立て／既存事業を拡充する道もある――そのうちの、もっとも現実的な策は、国立国会図書館の WARP 事業内で民間のウェブサイトの許諾を得ない形での制度的収集を可能にする方向に働きかけることだと考えるが、それは全く容易なことではない。

第三に、例えば、国際連合のような世界的な組織が、世界中すべてのウェブサイトをアーカイブすることも可能だろう。各国がそれに資金を供与し、全世界がまとまってそれを実施する。ウェブサイトを世界的な共有財と見なす考え方である。ただし、この試みそのものは超国家的なものであるとしても、その基盤になるのは各国の著作に関わる権利であり、やはり「社会」がどのようにウェブアーカイブを位置づけるのかという問題の外に出るものではないと言いうる。

このように日本社会がウェイバックマシーンをいかに位置付けるか、不足があるのだとすればどのように補おうとするのか。これは簡単に言いうることではない。本稿は、その議論の小さな土台を作った。とくにすべてのウェブサイトを許諾を得ず制度的かつ網羅的な収集によってアーカイブすることがいかに可能であるのか／ないのかについては、引き続き研究課題としていきたい。

■註

★01　本稿では「ウェブアーカイブ」、「ウェブアーカイブズ」、「ウェブアーカイビング」の 3 つの用語を区別して用いる。「ウェブアーカイブ」は、ウェブ上の資料のアーカイブそのものを概念的に名指すものである。「ウェブアーカイブズ」は、ウェブアーカイブによって集められた資料の総体を指す。「ウェブアーカイビング」は、ウェブアーカイブという「行為」の側面に着目した場合に用いる。

★02　ボーンデジタルとは、直訳すれば「デジタル上で生み出された（資料）」という意味だが、デジタル上で作成・流通し、デジタル上で消費されることを前提としたコンテンツを意味する。

より明確に言えば、紙などの「アナログ」媒体で利用されることを「想定していない」コンテンツのことである。

★03　IIPC は国際的なウェブアーカイブに関するコンソーシアムである。ウェブアーカイブに関する技術開発の国際協力などが行われており、日本からは国立国会図書館が加盟している。この立ち上げについてはインターネットアーカイブが深くかかわっており、柊ら［2008］は「IIPC は 2003 年の 7 月にアメリカの NPO である Internet Archive の呼びかけに応じた，欧米の国立図書館を中心とする 12 機関で発足した組織である」［2008:389］としている。日本からは国立国会図書館が 2008 年 4 月より加盟。

★04　収集そのものはクローリングによって網羅的に行うが、その公開は選択的にするという運用もなされる。ある範囲（例えば、公開の許諾を得た範囲）だけのアーカイブを作成するとしても、それはどのクローラの範囲を制約することに直結はしていないことに注意が必要である。

★05　通常のウェブページと異なり、Instagram や Tiktok のようなソーシャルネットワーキングサービスは、Heritrix の技術的制約からアーカイブが困難な場合がある（Schafer ら［2021:132-133］）。Schostag［2020］は、Twitter は Heritrix による収集に適していると報告している。

★06　インターネットアーカイブがどのようなクローリングポリシーによって、Heritrix をはじめとするクローラを運用しているかは定かでない。ここではあくまでも一般的なクローラの機能について述べている。

★07　Alex［2018］は、インターネットアーカイブの Heritrix に関する Github レポジトリに掲載された解説文書である。冒頭、「これは Heritrix 1.14.x の動作について説明していますが、ほとんどの場合は H2/H3 にも当てはまります」（Alex［2018］）と注記されている。

■文献

東 泰正　1999　「インターネットに関する著作権侵害とフェア・ユース原則の適用について」,『帝京短期大学紀要』(11):67-74

Osborne,Alex　2018 "Frontier queue budgets", Github(internetarchive/heritrix3),
URL: https://github.com/internetarchive/heritrix3/wiki/Frontier-queue-budgets

ブリュースター・ケール　i2011　「ブリュースター・ケール氏に聞く本の未来」,『マガジン航』聞き手：仲俣暁生,
URL: https://magazine-k.jp/2011/09/12/interview-with-brewster-kahle/

原田 隆史　2008　「Web アーカイブの仕組みと技術的な特徴（＜特集＞Web アーカイビングの現状と課題）」,『情報の科学と技術』58(8):383-388

Hardy, Q 2009 "The Big Deal: Brewster Kahle. Forbes",
URL: https://www.forbes.com/2009/11/25/alexa-amazon-entrepreneur-intelligent-technology-kahle-big.html

柊 和佑・阪口 哲男・杉本 重雄　2008　「世界の Web アーカイブ-IIPC（International

Internet Preservation Consortium）を中心にして（＜特集＞Web アーカイビングの現状と課題）」,『情報の科学と技術』58(8):389-393
Internet Archive 2019 "Form 990 for period ending December 2019",
URL: https://projects.propublica.org/nonprofits/organizations/943242767
Internet Archive 2017 "IMLS grant to advance web archiving in public libraries",
URL: https://blog.archive.org/2017/07/18/imls-grant-to-advance-web-archiving-in-public-libraries/
Internet Archive 2014 "Internet Archive's Terms of Use, Privacy Policy, and Copyright Policy",
URL: https://archive.org/about/terms.php
Internet Watch　2001　「過去 5 年間の 100 億ページもの Web ページを保管した Web アーカイブが公開」,『Internet Watch』,
URL: https://internet.watch.impress.co.jp/www/article/2001/1029/wayback.htm
国立国会図書館　2020　「International Internet Preservation Consortium (IIPC)――世界のウェブアーカイブ（おすすめコンテンツ）」, 国立国会図書館 HP,
URL: https://warp.da.ndl.go.jp/contents/reccommend/world_wa/world_wa01.html
国立国会図書館　2016　「国の機関サイトの残存率」,
URL: https://warp.da.ndl.go.jp/contents/reccommend/collection/linkrot.html
国立国会図書館　不詳　「令和 5 年度歳出予算額（当初）」,
URL: https://www.ndl.go.jp/jp/aboutus/outline/finances/budget.html
前田 直俊・大山 聡　2017　「ウェブアーカイブを支える技術」,『情報の科学と技術』67(2):73-78
Mohr, G., Stack, M., Rnitovic, I., Avery, D., Kimpton, M. 2004 "Introduction to heritrix", *In 4th International Web Archiving Workshop*:109-115,
URL:https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=7d4e01113bdb8958428a64bc07645444c01d062e
中井 良平　2022　「ウェブログアーカイブの必要性と課題」,『遡航』4:56-69
仲俣 暁生　2011　「あらゆる知識にユニバーサル・アクセスを」,『マガジン 航』,
URL: https://magazine-k.jp/2011/06/02/universal-access-to-all-knowledge/
Schafer, V, Winters, J　2021 "The values of web archives", *International Journal of Digital Humanities* 2(1-3):129-144
Schostag, S 2020 "The Danish coronavirus web collection – coronavirus on the curators' minds. ", *International Internet Preservation Consortium Blog*,
URL: https://netpreserveblog.wordpress.com/2020/07/29/the-danish-coronavirus-web-collection/
新保 史生　2008　「ウェブ・アーカイビングと法（＜特集＞Web アーカイビングの現状と課題）」,『情報の科学と技術』58(8):376-382

塩崎 亮　2019　「日本の大学ウェブサイトのアーカイブ状況――Internet Archive と WARP の比較」,『聖学院大学総合研究所 Newsletter』29(2):4-10
時実 象一　2016　「デジタル・アーカイブで世界をリードする Internet Archive 最近の動向」,『情報の科学と技術』66(9):490-494
山口 和紀　2022　「ウェブアーカイブの公開を支える法律と仕組み――社会運動のウェブアーカイブズ構築に向けて」,『遡航』4:70-86
山口 和紀　2023　「日本のウェブアーカイブはいかに形作られたか――1990 年代末から 2000 年代初頭にかけて議論を通して」,『遡航』7:37-58

■案内

本稿は、大幅に増補・改稿したのち、立命館大学生存学研究所より刊行予定の叢書の 1 冊に収録する予定である。詳しい情報は「叢書 身体×社会」（http://www.arsvi.com/ts/s.htm）にて随時更新する。

Collecting and Publicizing the World's Websites Along a Timeline

What is the Wayback Machine?

KAZUNORI Yamaguchi

Abstract:
The records on the web worldwide are currently disappearing. There is insufficient discussion on how to address this loss. Therefore, this paper focuses on the Wayback Machine, one of the pioneering web archiving systems developed by the Internet Archive, and examines its history, management, technology, and legal aspects from multiple perspectives. The Wayback Machine collects and publicly shares web information to prevent its loss, and this endeavor has achieved a certain level of success. Particularly notable is its attempt to archive all websites without requiring permissions, which is rare and valuable globally. However, there are technical, legal, and managerial challenges. How to perceive and position the Wayback Machine in society remains a future task.

Keyword:
Archive, Web Archive, Internet Archive, Wayback Machine