利用者によるウェブアーカイブの意義、技術とその課題

武田俊之 1)

1) 関西学院大学　高等教育推進センター

takeda@kwansei.ac.jp

Web Archiving by Users – Technologies and Challenges

Toshiyuki Takeda 1)

1) Center for the Studies of Higher Education, Kwansei Gakuin University.

概要

組織の改変、廃止、情報システムの変更によって、インターネット上のコンテンツが消失することがある。このような社会文化的な損失へ対処する方法の一つがウェブアーカイブの構築である。ウェブアーカイブの構築の技術コストは高価であったが、OSS として開発された新しいツールによって、個人であっても利用価値の高いアーカイブの作成が容易になっている。本稿ではこのような利用者個人のアーカイブ作成方法について説明をおこない、今後の意義や課題について述べる。

1　はじめに

インターネット上の情報はますます増加をつづけているが、一方で日々情報の削除がおこなわれている。学術情報に限っても、組織の改変、廃止、情報システムの変更によって、少し前に閲覧できた情報が見えなくなっていることも多い。消失した情報には、すぐれたコンテンツやその時点の社会状況を反映した有用なものも少なくない。

このような情報の消失、散逸による社会文化的な損失に対応するための一つの方法が、ウェブ情報を保存するウェブアーカイブである。ウェブアーカイブとは「ウェブサイトを定期的に収集し、累積的かつ長期的に保存して、元の状態のまま再現する事業」で、国立図書館等を中心に行われている[1]。1996 年に米国の Internet Archive がウェブアーカイブを開始して以来、2013 年には 120 以上の機関で実施されている。

日本においても、国立国会図書館のインターネット資料集保存事業（Web Archiving Project: WARP）が 2002 年から国内のウェブサイトを保存している[2]。公的機関が発信するインターネット情報については、平成 21 年 7 月 10 日に改正された国立国会図書館法に基づいて、国の機関、独立行政法人、国立大学法人、特殊法人等（以上 24 条）、地方公共団体、地方公社等（以上 24 条の 2）を定期的に訪問、網羅的に収集している（25 条の 3）。それ以外のウェブサイトについては、公益法人、私立大学、政党、国際的・文化的イベント、東日本大震災に関するウェブサイト、電子雑誌などを主な対象として、発信者の許諾を得られたものを収集・保存・提供している。アーカイブされたコンテンツは、21 年には 13,822 件（ファイル数 110 億個、2,387.753TB）までに登っている。

WARP 事業のコンテンツが充実する一方で、日本には大学や民間のウェブサイトで公開された情報を組織的に保存するアーカイブが存在しない。日本以外では、米国の非営利組織 Internet Archive（IA）の Wayback Machine[3]が、ウェブページを世界中から収集、公開するウェブアーカイブとして有名で実績がある。

個人が必要な民間のコンテンツについてはそれぞれが適宜適当なツールによって保存すればよい。しかし、規模の大きい体系的なコンテンツを個人が保存することは容易ではない。

たとえば、2022 年 9 月 30 日に組織が廃止された京都大学高等教育研究開発推進センターでは、サービス終了[1]が 8 月に公表という性急さであったが、ウェブサイトには数多くの有用な大学教育に関するコンテンツ―FD（Faculty Development）、オンライン授業、オープン教材など―が公開されており、サービス停止による知的資源の消失の影響

[1] 本稿執筆時点で、更新は停止されるが、当面維持されると広報されている（2022/9/30 付）http://www.highedu.kyoto-u.ac.jp/news/news-1716/

は大きい。このような組織の廃止等によるコンテンツの喪失は、大学だけではなく学会やNPO等でもおこりうるが、個人や民間によるウェブアーカイブ構築が容易になり、アーカイブ連携を実行できれば知的資源消失の影響を少なくすることが期待される。

この論文では、ウェブサイトを利用する個人がアーカイブを構築するための技術を紹介するとともに、社会的な意義と今後の課題について述べる。

2 ウェブアーカイブの技術と手順

2.1 ウェブアーカイブとは

先に述べたように、本稿ではウェブアーカイブを定期的収集、累積的かつ長期的保存、元の状態のままの再現が可能なものとする。検索エンジンやデータ分析用のデータ収集は、ウェブアーカイブと一部技術を共有するが、目的や結果が異なる。

検索エンジン等のページ収集がウェブアーカイブと最も異なる点は、ウェブページの再現を目的としていないところである。収集対象は主にデータとなるページであり、表示に関連するCSSやJavaScriptなどは重視されない。また、収集したページから必要な情報だけが取得されて（スクレイピング）加工されることが多い。一方のウェブアーカイブはブラウザ上での表現が同一になることを目標としてページが収集される一方で、データとしての内容には関心がない。また、HTTPリクエストやレスポンスのデータも再現表示や管理のために累積的かつ長期的に保存される。

2.2 ウェブアーカイブ化のプロセス

(1) 選定

アーカイブ化するウェブサイトを選定する。国会図書館のように法的根拠を持って収集する場合はバルク収集（bulk harvesting）をおこなうこともあるが、利用者がアーカイブを作成する場合には特定のサイトの選択収集（selective harvesting）をおこなう。

(2) クローリング

クローリング（crawling）とは、ウェブページを巡回、収集することである。多くの場合、ロボットによるクローリングがおこなわれるが、本稿では後に手動による巡回方法も紹介する。

クローリングではウェブサイトのトップページからリンクを順次たどってページを収集するか、収集ロボット向けのサイトマップに指定されたURLにアクセスして、コンテンツのURLを取得する。しかし孤立した（orphaned）ページの存在や、サイトマップにサイト内のすべてのページが記述されているとはかぎらないことに注意が必要である。

(3) 保存

クローリングした結果を、レスポンスだけではなく、アクセス時のリクエストやメタデータを含めてアーカイブに保存する。

クローリングの頻度については、対象となるウェブサイトの更新頻度や負荷、アーカイブの容量などを含めて決定する。

ツールによっては、ページの更新がない場合は収集をおこなわない差分収集の機能がある。

(4) 組織化

収集時にウェブページの情報（URL、リクエスト、レスポンス、コンテンツなど）を用いてインデックスやメタデータを付与する。メタデータは人手によって付与される場合もあり、国会図書館はDC-NDLに準拠記述規則によってメタデータを作成している[1]。

後の検索のために、ページ保存時に全文検索用など、後付のインデックスが付与される場合もある。

(5) 利用

ウェブページを再現する（リプレイ）システムのよって、URLと取得時刻等の情報によってアーカイブから取り出したページを利用する。

公的なウェブアーカイブでは公開の範囲と方法を選択する必要がある。

2.3 ウェブアーカイブの保存形式

この節では、ページの再現や累積的かつ長期的保存を重視したウェブアーカイブの保存形式について説明する。

HTMLで表現されたウェブページのアーカイブ化は、単にそのHTMLファイルを保存すればよいわけではない。ページの表示は、CSSやJavaScriptを読み込んでレンダリングされるものであり、関連するファイルを含めて、保存する必要がある。DOM（Document Object Model）の操作や、非同期で他のコンテンツを読み込んでおこなうAJAXなどの動的レンダリングの配慮も必要である。

ウェブページの保存用のフォーマットとして、MHTML（MIME encapsulation of aggregate HTML documents、拡張子は多くの場合.mht）とWARC形式が用いられる。MHTMLは主に個人な

どが単一のウェブページを保存するために用いられ、WARC 形式はウェブアーカイビングの標準フォーマットである。ここでは WARC 形式について説明する。

WARC 形式は Internet Archive が採用していた ARC 形式を拡張したフォーマットで、2009 年 5 月に ISO 28500:2009 として標準化された（現在は ISO 28500:2017）[4]。

WARC 形式は、1 つ以上の WARC レコードで構成される。WARC レコードはヘッダーとコンテンツブロックで構成される。図 1 において、空行より前が WARC ヘッダー、それ以後がコンテンツブロックである。

ある URL に HTTP プロトコルでアクセスしたとき、WARC レコードタイプが request と response の WARC レコードが作成される。このときのコンテンツブロックは、それぞれ HTTP プロトコルのリクエスト情報とレスポンス情報そのものである。

WARC レコードタイプは他に、warcinfo、resource、metadata、revisit、conversion、continuation があり、コンテンツ以外にアクセス時の情報やメタデータを保存することによって、データ履歴やデータ交換などウェブアーカイビングに有用な情報を記録することが特徴である。

図 1 WARC レコードの例

```
WARC/1.0
WARC-Type: response
WARC-Record-ID: <urn:uuid:6153a256-446a-11ed-907f-12207609c342>
WARC-Target-URI: https://www.kantei.go.jp/
WARC-Date: 2022-10-05T04:58:40Z
WARC-IP-Address: 18.65.159.78
Content-Type: application/http; msgtype=response
Content-Length: 34702
WARC-Payload-Digest: sha1:RJ77WE7HPP3ZT6J4S7HHEO7QPMZTHDC2
WARC-Block-Digest: sha1:HLB5LQI4GM4T2WHV7IWM53L4TP7V7OWF

HTTP/1.1 200 OK
Content-Type: text/html
Content-Length: 33868
Connection: keep-alive
x-amz-id-2: H0oPDkpFhnlvcaG4nh8qtN6cwqt0fiKSa6WXSoUI2R7uPxhh7oCv
x-amz-request-id: 9K4SF3ZXPFJHXTS3
x-amz-replication-status: COMPLETED
Last-Modified: Wed, 05 Oct 2022 02:46:27 GMT
x-amz-server-side-encryption: AES256
x-amz-meta-user-agent: AWSTransfer
x-amz-meta-user-agent-id: kantei-prd-apn01-sftpuser-kantei01-ala
x-amz-version-id: G9G5ejQuhMmiLJxLwisGWoCozIs522rx
Accept-Ranges: bytes
Server: none
Date: Wed, 05 Oct 2022 04:58:41 GMT
ETag: "702c6c0104c9546eccfdc128a5a4d6ab"
X-Cache: RefreshHit from cloudfront
Via: 1.1 f28de56dcc4be3921b3badb7d47b0b10.cloudfront.net (CloudF
X-Amz-Cf-Pop: NRT51-P2
X-Amz-Cf-Id: lGZuTcnOWnY_sZveN63oehiA0b8I3E6CM0vUPPyFh21iJd62v9V


<!DOCTYPE html>
<html lang="ja">
<head>
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta name="viewport" content="width=device-width, initial-scale
<meta name="description" content="首相官邸のホームページです。内
<meta name="keywords" content="首相官邸,政府,内閣,総理,内閣官房"
<meta property="og:title" content="首相官邸ホームページ">
```

技術的制限

ウェブコンテンツの性質から、以下のようなコンテンツの正確なアーカイブ化は容易ではない。

1. バックエンドにデータベースがあるウェブサイトの検索等の動的コンテンツ
2. 乱数等を利用したゲームなど、アクセスするたびに内容が変わるコンテンツ
3. ストリーミングと連携したコンテンツ

ウェブ技術の進展によりさらに変化も予想されるため、制約を配慮しつつ、目的の範囲でウェブアーカイブをおこなうことになる。

3 アーカイブツールとその利用例

3.1 ウェブアーカイブのツール

ウェブアーカイブは、国会図書館や Internet Archive のように網羅的におこなう大規模な実施、大学や NPO 等の組織が目的をもって行う中小規模の実施、MHTML 形式で保存するような個人規模の実施に分類される。

大規模なアーカイブに使用するツールは、プロセスのステップごとに専用のツールを用いることが多い。国会図書館では、クローラーと保存に Heritrix[5]、組織化に NutchWAX[6]や Solr[7]、リプレイに OpenWayback[8]を使用している[1]。Heritrix や OpenWayback はウェブアーカイブ専用であり、実施に必要な技術レベルは高い。

中小規模では、Archive-It や Conifer などのウェブアーカイブ機能を提供する SaaS が用いられることが多い。この方法は運用コストを下げることができるが、利用面での制約が大きい。MHTML の保存は系統だったアーカイブの作成や利用面で制約がある。これらの機能面の制約を低減しつつ、比較的低コストに柔軟なアーカイブの運用を可能にするために、Webrecorder プロジェクト[9]がツールを開発している。

3.2 Webrecorder プロジェクトについて

Webrecorder プロジェクトは”Web archiving for all!”という標語を掲げて、オープン・ソースの Web アーカイブ・ツールを提供している。ツールには、キャプチャ、リプレイ、アーカイブの操作、クローリング等 Web アーカイブに必要なものが含まれている。

ツールの中でも pywb[10]はウェブアーカイブ化を支援するツールで、ウェブページのキャプチャからリプレイまでを Python でコーディングす

ることができる。warcio[11]は pywb の基礎となる WARC ファイルの操作ライブラリである。

3.3 ウェブレコーディング

pywb をインストールして、アーカイブの初期化が終わると、ウェブレコーディングが可能になる。” wayback --record --live -a --auto-interval 10” を実行すると、ブラウザで
”http://localhost:8080/record/<collection>/archive/record/http://example.com/ “
にアクセスするだけで、http://example.com がアーカイブされる。画面上のリンクをクリックすることで、リンク先のページも表示、保存される。

3.4 クローリング

Webrecorder プロジェクトではクローリングをおこなう Docker コンテナ Browsertrix Crawler も提供している。高機能で利便性も高いツールであるが、エラー発生時のリカバリー制御がむずかしい。クローリングには、サイトマップの利用（robots.txt に記述されたサイトマップから、アーカイブする URL を収集）か、スクレイピング（ページをパースして、リンクしている URL を取り出す。対象となるサイトの配下ではない URL はフィルターする）の方法がある。

3.5 ウェブページのリプレイ

ウェブページのリプレイ（再現表示）は ”http://localhost:8080/<collection>/archive/record/http://example.com/ “ へのアクセスで可能である。

3.6 ウェブアーカイブの作成例

京都大学高等教育推進センターおよび京都大学 OCW のウェブアーカイブを作成した。これらのページの多くは WARP でも保存されておらず、Wayback Machine でも 7 割程度の保存である。

それぞれのサイトにはサイトマップが存在したが、センターの方にはコンテンツの一部しか含まれておらず、OCW の方（WordPress が自動的に作成）はコース内の URL と体系が異なったものがあり、サイトマップとウェブページのリンクを抽出するクローラーを併用して、pywb レコーダーにアクセスするスクリプトを記述した。

4 まとめと今後の課題

以上、Webrecorder のようなすぐれたツールを使うことによって、ウェブサイトの利用者個人であっても、アーカイブ作成やアーカイブを利用するコードを書くことが比較的容易に可能であることを示した。以下では利用者によるアーカイブ作成の意義と今後の課題について述べる。

組織やサイトの廃止、改変が予告されることはまれであり、ウェブ情報の消失の多くは気づかれない。個人が WARC 形式で情報を保存することは、事後にページ（サイト）の再生を可能とする。

しかし、アーカイブの公開には、著作権や個人情報等の問題がある。今回対象としている大学や学術情報のサイトについては、Creative Commons 等の自由な利用を許諾するライセンスを付与する慣習が望まれる。

技術面では発展をつづけるウェブ技術への対応である。先に述べたように動的なページのアーカイブには技術的な課題がある。

ウェブアーカイブの今後の最も重要な課題は利用法の開発であり、技術面では個人アーカイブ間の連携であろう。個人アーカイブは大規模アーカイブと比較して収集が完全ではないが、個人の関心を反映した対象の収集には価値がある。アーカイブの統合的な検索、閲覧が可能にするために、アーカイブの連携のために開発された Memento プロトコルが開発されている。このような広範に連携したアーカイブの実現は、人文学などを含めた新しい研究につながることが期待される。

謝辞

本研究の一部は JSPS 科研費 20H01713 の助成を受けたものです。

参考文献

[1] 前田直俊, 大山聡, ウェブアーカイブを支える技術, 情報の科学と技術 67 巻 2 号, 73-78, 2017
[2] 国立国会図書館, インターネット資料保存事, https://warp.da.ndl.go.jp/
[3] Wayback Machine, https://archive.org/
[4] ISO 28500:2017 Information and documentation — WARC file format, 2017 https://www.iso.org/standard/68004.html
[5] Heritrix, https://github.com/internetarchive/heritrix3
[6] NutchWAX, https://archive-access.sourceforge.net/
[7] Solr, https://solr.apache.org/
[8] OpenWayback, https://netpreserve.org/web-archiving/openwayback/
[9] Webrecorder, https://webrecorder.net/
[10] Pywb, https://github.com/webrecorder/pywb
[11] Warcio, https://github.com/webrecorder/warcio