ウェブログアーカイブの必要性と課題

立命館大学大学院先端総合学術研究科

中井 良平

要旨：
本稿では、ウェブログ（以下：ブログ）アーカイブが有する意義を概観した上で、その方法と課題について検討した。その結果、以下のことが明らかとなった。①ブログ記事上の語りを分析した国内外の先行研究が、ブログ記事の特性をどのように考察しているかを見ていくことで、先行研究において、ブログ上の当事者の語りに固有の価値が置かれていることがわかった。②筆者らのウェブスクレイピングの取り組みから、大規模なブログアーカイブが、技術的には個人が用意できる環境で十分に可能なものであることが明らかになった。③同様の取り組みとその考察から、残る法的な課題等も、関係各所がデジタル情報のアーカイブに関する議論を行なっていく中で、十分に解決可能なものであると結論した。

キーワード：
デジタル情報のアーカイブ、ウェブログ（ブログ）、ウェブスクレイピング

0. はじめに

2000 年代以降にインターネット環境の普及と同時に急速に人々の間に広まったウェブログ（以下：ブログ）は、今なお多くの新規記事が書かれ、また過去に書かれた膨大な数の記事が存在する。

本稿でも紹介するように、ブログ上には多くの人々の語りや記録が存在し、そこには病の語り、戦争・災害の被災の語り、公的な活動の記録など、積極的にアーカイブが行われるべき対象が無数に含まれている。

一方で、ウェブ環境の変化により、2019 年の「Yahoo!ブログ」を筆頭に、大手のブログサービスの終了が相次いでいる。

しかしながら、SNS(ソーシャルネットワーキングサービス)のデータを含む、個人が生み出したデジタル情報をどのようにアーカイブしていくのか、という議論は、学術領域において端緒についたばかりで、とりわけ日本ではその議論が殆どなされていないようである。

現代においては、私的／公的な領域を問わず、人が生み出す情報の殆どはデジタルなものになっている。そして、適切な保存やバックアップが行われなければ、デジタルデータの寿命は、紙に書かれた情報よりもはるかに短い。この事実は、デジタルデータのアーカイブについての議論・方法の確立が遅れれば遅れるほど、従来とは比較にならない速度で、人々が生み出した情報

が失われていくことを意味している。

技術的には、大きな機関ではなく個人の研究者であっても、大量のデータをアーカイブする土壌は整っており、具体的な課題をクリアしていく段階にきている。筆者らのブログアーカイブの取り組みを例にそのことを解説していく。

1. 日本におけるブログサービスとアーカイブの動きの現状

本節では、日本におけるブログサービス終了の動きについて、及び、日本におけるブログアーカイブの現状を簡単に見ていきたい。

まず、以下は過去に終了した／終了がアナウンスされている主なブログの一覧である。既にサービスが存在せず、得られる情報が限られていることから、オフィシャルな情報に限らず、当時の報道や、利用ユーザーが残しているブログも参照した。また、各ブログのユーザー数に関する情報は公開されていない場合が多く、運営会社の知名度などから掲載を判断した。

終了した主なブログサービス			
サービス名	運営	開始	終了
CURURU	現 LAIN	2005/05	2010/03★01
Windows Live Spaces	Microsoft	2004/08★02	2011/03★03
au one ブログ	KDDI	2005/05	2011/03★04
Maglog	ベクター	2006/10★05	2012/05★06
OCN ブログ人	NTT コミュニケーションズ	2004/03	2014/05★07
Yahoo!ブログ	Yahoo! JAPAN	2005/01	2019/12★08
ヤプログ！	GMO メディア	2004/06	2020/01★09
CROOZ blog	クルーズ	2005/07★10	2022/05★11
AutoPage	GMO メディア	2004/07★12	2022/08★13
ウェブリブログ	ビッグローブ	2004/03	2023/01★14

表からは、2010 年代の前半に一度目のサービス終了の波があり、2019 年から二度目の波が起こっていることがわかる。中でも 2019 年終了の Yahoo ブログは、知名度・ユーザー数で最大手の一つだったと考えられる。そのため、同ブログの終了時にはユーザーに混乱が生じ、自らを「Yahoo!ブログ難民」と称する人たちが多く生まれた。同ブログでは、データのインポート／

エクスポート機能が用意されていなかったり★15、Yahoo! JAPAN が用意したデータ移行ツールを用いてもユーザー間のコメントデータの一部しか移行できなかったり★16 と、データ移行に大きな制限があったとのことである。

他方、公的に行われているブログアーカイブで今回唯一発見できたのは、国立国会図書館と株式会社サイバーエージェント社（以下：サ社）の協力で行われているという、「アメーバブログ」の一部タイトルの保存である。筆者が国立国会図書館インターネット資料収集保存事業（WARP）担当課にメールで問い合わせたところ、2022 年 10 月 26 日現在（以下同）、国立国会図書館に収められているブログのタイトル数は 42 件とのことであった。また、ブログを対象とした収集方針が定まっていないために、積極的な収集は行なっておらず、受け入れの希望に対しても、すぐに対応できる状況ではない、とのことであった。サ社との協力の経緯であるが、サ社から WARP に対し収集の打診があり、権利処理をサ社側で一括して行うという両者の合意のもと、収集が開始された、とのことである。サ社の発表によれば、2019 年時点でアメーバブログを含む「メディアサービス」である「Ameba」の会員数は 2019 年時点で 6000 万人とのことであり★17、そのうちのどの程度の人がブログを所有しているかは不明であるが、42 件という数字は、極めて例外的なものであることがわかる。

また「ウェブリブログ」の運営元である BIGLOBE に問い合わせを行ったところ、サービス終了後のアーカイブの予定はないとのことであった。今年 5 月に終了した CROOZ blog も「17 年間で延べ 500 万人の皆様による 30 億件」の記事を、「サービス終了後に全て消去」予定とのアナウンスが行われていた★18。ブログサービスの終了後、そのデータは消去される運命にあるらしい。

今後もブログサービスの終了は続いていくと考えられるが、現状においては、アーカイブの試みは実質的に存在しないと考えられる。

2. 先行研究におけるブログ

多くの場合、ブログは専門家が書いたものではない。そこには日々の出来事の感想といった、アーカイブする価値をわずかでも有しているのかどうか、一見してわからない情報が無数に含まれている。もちろん、そのような情報も含めアーカイブしていくことが重要であるという議論はあり得る。選別を経てアーカイブされたデータには、選別を行った者の関心が反映されており、それゆえにそのデータの利用にも限界が生じる。逆を言えば、未選別のデータ群には、研究者の予測を超えた価値が内包されている、と言える。また、膨大なデータの中からどのように必要な情報を探し出して選別するのかという問題もある。技術的に可能であるならば、まずはデータ全体のアーカイブを行い、その後それぞれが利用法を考えていく、という順序が、おそらく正しいものであろう。

しかし、アーカイブされたデータがどのような研究の可能性を有しているのかが例示されなけ

れば、ブログ記事のように現在はほとんど関心が払われていないデータのアーカイブの必要性について、理解されることは難しいかもしれない。

そこで本節では、ブログ記事を分析対象としたいくつかの先行研究を紹介することで、その例示を行おうと思う。

①Gloria 他［2012］はイラク戦争時にイラク人によって書かれたブログを分析したものであるが、WEB 上の SNS を、大規模な災害が人々へ与える影響を長期的に調査する機会を研究者に提供するものだと評価する。Gloria らは、ブログの特徴として、紛争地域など外部の人間が立ち入ることのできない場所の情報を提供することや、偏向的な既存のメディアが報道しない当事者の視点を伝えることを挙げている★19。

②Luisa 他［2019］は 50 歳未満の乳がん患者のブログの分析を行なったものであるが、病者自身による病に関するブログを、「オンラインの病の語り」と形容している。また、がんに関する女性たちのウェブサイトを（サイバー）エージェンシーの一形態と見做すという文脈で、（現在の）技術装置以前の伝統的な病の語りにおいては不可能だった、本物の病の語りの忠実な記述が可能となった、という McNamar の言及が引用されるなどしている★20。

③坂口［2012］は、自傷行為を行う中学・高校生のブログを分析したものである。坂口は、自傷行為を行う未成年へのインタビュー調査は倫理的に困難だとし、ブログを採用した★21。親や教師、友人に語ることができず隠さなければならない、自傷についてや自身の気持ちを、「吐露する場」がブログ空間であると捉えている★22。

④Kyoung & Yi-Fan［2019］は、ブログの内容を分析した前掲の 3 つとは異なり、HPV ワクチンを推奨する立場から、「ワクチンを接種しがんを防げた」／「打たなかったためにがんになった」という文脈のブログを読んだ読者の反応の分析を行なっている。そして、分析結果から、SNS を利用した「有効な」ワクチン推奨キャンペーンを考察する★23。

①②は、伝統的なメディアが伝えなかった当事者の声を、ブログが伝えると評価している。①③は、立ち入ることが危険な地域にいる人々や、調査の実施に倫理的に懸念が伴う人々の声を、ブログが伝えるとする。①は、災害が人々に与えた影響を長いスパンで検証できるとする。②はブログ空間が、オンライン上の当事者のコミュニティであるとし、その重要性を指摘する。

以上のように、総じて、先行研究においては人々が書いたブログを語りの記録とみなしていた。付け加えるならば、長期に渡り、当事者間の交流もなされながら書かれたブログ上の語りは、部外者としての研究者が短いスパン・そう多くない回数行なったインタビューに現れる語りとは異なるものになっていると考えられる。そのことがブログ上の語りの記録のアーカイブに、社会的な・研究上の意義を与えることは言うまでもないだろう。

また④のような研究の存在は、ある出来事に関する人々の語りが、それを読んだ人々に影響を与えたり、キャンペーン等に利用され得るものであることを示唆している。網羅的なブログデータのアーカイブと、長いスパンでのデータの分析で、多くの人にとっての関心の的であった出来事に関する言説がどのように形成されて・変遷していったのかが明らかにされる可能性がある。

3. ブログサービスの終了で失われるものと人による作業の限界

次に、あるブログサービスの終了により、どのような語り・記録が失われるのかを、あるブログ（以下：A ブログ）に対する筆者らのアーカイブの試みから得られたデータにより示したい。

ブログの種類（いずれもブログ紹介文から分類）	
障害当事者団体によるブログ	当事者形
第二次対戦経験者によるブログ	当事者形
闘病に関するブログ	当事者形／医療形
医師によるブログ	医療形
医療関係者によるブログ	医療形
病院経営者によるブログ	医療形
医療機関によるブログ	医療形／公系
議員の公式ブログ	公形
介護職によるブログ	福祉形
新潟県中越大震災の記録	災害記録
東日本大震災の記録	災害記録
市民オンブズマンによるブログ	運動形
NPO 団体のブログ	運動形
反戦運動家によるブログ	運動形
労働組合によるブログ	運動形
大学教員及び研究者によるブログ	研究形
排外主義思想のブログ	言説形
ジャーナリストによるブログ	言説形

上記リストは、作業開始当初に、筆者の関心によるキーワード検索、及びその検索結果から想起されたキーワードでの検索によるものである。作業時間は 3 時間程度であり、系統だったものでも、網羅的なものでも全くない。また、分類も極めて大まかなものとなっている。

このうち、筆者の専門領域との関わりで、2022 年 9 月下旬から 11 月にかけてリスト化されたのが、闘病に関するブログ（以下：闘病ブログ）であり、約 400 件となっている。方法としては、Google 検索（以下：G 検索）で検索コマンド「site:」を用いて、検索範囲を A ブログ内に限定★24 ののち、厚生労働省が作成した指定難病一覧表などを参照しながら、病名による検索を行った。作業時間は一回 1〜3 時間程度で、約 1 カ月の間、断続的に行った。また検索の過程で発見された多くの数を有する分類は、市議会議員を中心とする、議員の「公式」なブログである。

この甚だ未完成のリストは、アーカイブの根拠としてこれらのブログの存在が重要であるとい

うことではなく、重要であると思われるものを含みながらも、人の手による作業ではブログのアーカイブを進めるには限界がある、という事実を示していると考える。

前述のように、筆者は約 1 カ月という期間、少なくない時間をアーカイブ対象ブログの選定作業に充ててきたが、闘病ブログにその範囲を絞っても、400 本が集まったのみだった。また、当初は闘病ブログ以外のものについてもリスト化を行う予定であったが、闘病ブログのリスト化が思うように進まず、前述の、開始当初に行った作業以上を行うことはできなかった。

その大きな要因のひとつは、作業が検索サービスの機能に依存しているという点にある。検索サービスでは、（特に大まかなキーワードでの検索の場合）必ずしもキーワードと一致した結果が表示されるわけではなく、また、検索結果が多数だった場合、参照できるのはその一部であり、残りは表示されない。このことから、リスト化はもちろん、検索対象全体数の把握も困難となる。例えば、「闘病」と検索しても、表示されるのは検索上位の限られたページのみで、その他のページについては、別のキーワードでの検索を繰り返し、上位に表示されるのを待たなければ、その存在すら把握できない。

この点においても、網羅的にブログデータを収集する意義がある。すなわち、収集したデータをデータベース化することで、検索サービスでは大きな制限がかかった状態でしかできなかったサーチが、データ全体に対して行えるようになる。

次節では、網羅的なデータ収集の方法と課題について見ていく。

4. 網羅的ブログデータ収集の方法と課題

4-1. 方法

大量のデータを収集する場合、「ウェブスクレイピング（以下：スクレイピング）」という技術が用いられる。これは、Python などのプログラミング言語を用い、対象サーバーにアクセスし、指定されたデータ群を自動収集する技術である。収集データの設定は自由なカスタマイズが可能であり、例えばブログ上のデータであれば、タイトル、日付、本文、コメント、画像の一部や全部を取得することができる。データ収集に必要となるのは、対象となる HTML ページの URL であり、外部ファイルの URL リストを読み込むことも可能であるため、膨大な数のデータ収集を、短時間の操作で可能にする（※プログラムがデータを自動収集する時間を除く）。

ウェブサイト全体のデータを収集する場合、データ量が膨大なものになる恐れがあるが、テキスト情報（タイトル、本文……）のみを収集する場合、例えばあるブログサービス上の全てのデータを収集することが、個人が容易に準備できる環境で可能になる。また、スクレイピングのプログラム自体も、一般的な構造のブログからデータを収集するのであれば、さほど複雑なものではなく、プログラムは標準的な PC で動作する★25。つまり、一見すると大規模な計画であるように思われる大規模なブログデータのアーカイブは、個人でも十分に可能なものである、という

ことだ。必要な環境や時間については、考察で後述している。

課題としては、いかに全対象 URL のリストを作成するかという技術的な部分、及び法的な部分となってくる。しかしながら、事項で述べるように、これらは解決方法が見当たらない課題ではなく、各所の協力により十分に対処することが可能な課題であると考える。

4-2. URL リストの作成

前述の通り、あるサービスのブログデータを網羅的に集めるには、各ブログページの URL が必要である。その理由を簡単に説明する。スクレイピングプログラムは、人が手動でデータを集める際のそれに近い動きをする。例えば、次のようなものだ。まずブログの最新記事から本文など必要なデータを取得する。次に、ほぼ全てのブログで一般的と考えられる「ひとつ前の記事」へリンクされたボタンを探し出し、前の記事のページを開き再びデータを取得する。これが延々と繰り返される。そして、「ひとつ前の記事」ボタンが見当たらなくなると（つまり最後の記事ページまで行くと)、そのブログの記事の収集を終了し、外部のリストから新たな URL を読み込み、新たな収集を開始する。収集されたデータは、エクセルなどの外部ファイルに自動で書き込まれていく。このようにデータ収集が自動で行われる。この際必要な URL とは、最新の記事 URL（あるいはそこへリンクされたブログトップページの URL など）ということになる。

どのようにその一覧を作ればいいだろうか。多くの場合、ブログサービス内の各ブログには独自のサブドメインが割り当てられている。もしブログサービス提供者がサブドメイン情報を外部から取得できる状態にしていれば、サーバーにアクセスすることでその一覧を取得できるが、多くの場合、その情報は取得できない設定にされている。おそらく、この情報はサービスの運営戦略とかかわり公開されない。また、公開されている著作物とはいえ、その複製に必要な情報を各運営者から共有されるためには、デジタル情報のアーカイブに関する議論が不足している。

別の方法は、絶え間ない巡回を行い、網羅的にサーバー・サイトにアクセスすることで、巨大なデータベースを作成・一般に公開しているサービス、つまりウェブ検索システムを利用することだ。このシステムは完全なものではないが、日夜改善が行われているであろう同システム以上に目的に適うものは、おそらく存在しない。そして、おそらく G 検索が最も安定した検索結果を得られる。

しかし、ここでもまた、前項で見た、検索結果の非公開という問題がネックになってくる。そこで筆者は Google 社（以下：G 社）に対し、目的とするデータを有料で購入することができないかを問い合わせた。問い合わせは 2022 年 10 月 20 日頃に行い、11 月 6 日現在回答待ちである。G 社の回答次第ではブログデータのアーカイブが大きく前進する可能性がある。

4-2. 法的課題

各ブログ記事の著作権は、書き手であるサービス利用者一人一人にあると考えられる。本文を

引用したり、データを保存したりする場合は、著作権法に則る必要があるが、その他の著作物と同様、著作者の許諾がなくとも、限られた範囲での複製及び共有は可能なものと思われる。具体的には、研究者個人が、ダウンロードしたデータをオフラインで保管することや、少人数の研究グループで共有することは可能だと考えられる。アーカイブされたデータに公共的な価値がある場合、どのように課題をクリアしていくのかという議論が行われるべきであるが、ここではサービスの終了に伴うデータ消失という危機に際した、緊急避難的な保存活動に焦点を当て、議論を進めていく。

なお、得られたデータの研究利用に際しては、本稿で参照した先行研究の全てで、匿名化を行っていた。すでに公開が終了したデータを研究利用する場合、扱いにはさらに慎重になる必要があることは言うまでもない。

おそらく、データの保存そのものに関わる最も大きな法的課題は、大量のデータをダウンロードするためのサーバーへのアクセスをめぐる問題である。大量のアクセスに際し問題となってくるのが、ブログデータが置かれたサーバーへの負荷である。プログラムを用いたサイトへのアクセス・巡回は、現在一般的に行われていることであり、例えば、G 社などの検索サービス提供者や、ウェブ上の資料のアーカイブを行なっている国会図書館が、通常の業務として行なっている。しかしながら、ウェブ上でのデータの掲示を行なっている者が明示的にスクレイピングを禁じている場合や、サーバーへの負荷を理由にスクレイピングを行ったものが罪に問われる場合がある。

2010 年、落ち度のないスクレイピングを行った男性が、アクセス先サイト（岡崎市立中央図書館のサイト）に障害を引き起こしたとする図書館側の被害届により逮捕された（岡崎市立中央図書館事件）。後に、同障害はサイトの欠陥に起因して起こったものであったことが明らかになり、図書館側は男性に謝罪を行った。日本図書館協会に置かれた「図書館の自由委員会」は、岡崎市立中央図書に訪問調査を行い、「図書館が利用者の基本的人権の中でも中核である身体的自由を奪う結果になったことは残念であるが、和解に至った関係者の努力は大としたい」と報告している★26。

この事例から分かることは、プログラムを用いて適正なアクセスが行われた場合でも、相手サイトやサーバーの管理者の訴えに基づき、プログラム利用者が罪に問われ得る、ということだ。このような事例は、法やその運用の不備という文脈で解釈されるべきだと考えられるが、仮に個人が罪に問われてしまった場合、身体的拘束を受けるなど、その不利益は極めて大きく、スクレイピングを行おうとする者は萎縮し二の足を踏んでしまうことになる。現時点でリスク回避のためにデータ収集者が行えるのは、サービス提供者に許可を取った上で、後述のように、適切な頻度でのアクセス感覚でスクレイピングを行う、といったこと程度だろうか。今回筆者らは A ブログの運営元にスクレイピングの許可を得た上で、アーカイブの試みを行っている。

しかしながら、前述したように、プログラムによるサイト巡回は、多くの企業・機関が日常的に行っており、それら実施者が、予測不可能な訴訟リスクを抱えながら、極めて膨大な回数に及ぶアクセスを繰り返しているとは考えにくい。それら実施者は何かしらの方法・対策を有してい

るはずである。

5. 考察

リストが作成できた場合の、ブログアーカイブにかかる時間、手間、必要な記憶媒体の容量について考察を行う。現在アーカイブを検討している、A ブログを例に示す。

まず、アーカイブ全体に必要な時間であるが、前述の「site:」コマンドを用いて G 検索を行うと、A ブログのドメインには約 400 万ページが存在すると推測される（11 月 7 日時点、以下同）。また、別のコマンドを用いて検索を行った結果、ブログタイトルは約 28 万件と推測される★27。網羅的なアーカイブとは、このページ全てにアクセスし、ブログ本文などの情報を取得することを意味する。各ページへのアクセス間隔（＝相手サーバーへのアクセス間隔＝リクエスト間隔）はスクレイピングプログラム上で自由に設定できるが、前述の通りサーバーへの負荷を考慮しなければ罪に問われる可能性がある。ここでは、国立国会図書館などが採用している、1 秒 1 回という穏当なリクエスト間隔★28 を採用し話を進める。

これは 1 秒間にひとつのページから情報を収集することを意味し、つまり全ページからの情報収集には 400 万秒＝約 1111 時間が必要となる。これは一見膨大な時間に思えるが、プログラムを実行している間、実行者は一切の作業をする必要はなく、実行中の PC への負荷もそう高くはないため、適切な環境で継続して PC を動作・プログラムを実行するならば、実行者の作業量・時間は全体に比べごく短いものとなる。人の手が必要となるのは、プログラムの作成時と、不具合が起きた場合等の修正時であり、前述の通り、必要とされるプログラミング言語の知識をある程度有する者であれば、一般的なブログのスクレイピングに求められる技能はそう高度ではないと思われる★29。

相手サーバーへの負荷という観点から、同時に 2 台以上の PC から同一サーバーへのアクセスを行えないため、あくまでリクエスト間隔により、ひとつのブログサービス全体の収集に必要な時間は変わってくる。しかし、例えば、日本語のブログ全てをアーカイブする、ということが目指され、各自がブログサービスひとつ（「アメーバブログ」、「はてなブログ」…）を受け持った場合、最もページ数の多いブログの収集が終わった際には、他のブログの収集も全て終わっている、という計算になる。

おそらく、最もページ数の多い日本語のブログはアメーバブログだと思われるが、前述の方法で、ページ数は約 2540 万ページと推定される。2540 万秒＝7055 時間であり、単純計算では約 294 日、ということになる。実際にはさまざまな要因がありプログラムを全く休みなく動かし続けることはできないと思われるが、2000 年代前半からの日本中のブログを収集するとして、それは技術的には、何年もの時間が必要とされるような遠大な計画ではないということだ。そして必要とされる時間のほとんどは、プログラムが自動でデータを収集する時間である。

また、データ収集に際し必要とされる記憶媒体の容量についての考察は次の通りである。ここ

では画像を収集しない場合を想定する。筆者が A ブログ上のある闘病記のスクレイピングを行ったところ、記事数は約 430 件で、データを書き出したエクセルファイルの容量は約 500KB だった。簡易的に、前述の通り推測されている 28 万件のブログひとつひとつにその容量が必要だと計算して、１億 4000 万バイト＝140GB となり、現在流通している記録メディアの容量からすれば、取るに足らないものとなる。実際は、各ブログの平均記事数はこれより遥かに少ないはずであるから、必要な容量はさらに小さくなるだろう。おそらくテキストデータだけであれば、日本最大のアメーバブログの全記事であっても、小さな外部記録メディアに収まってしまう、ということになるだろう。またそのデータの収集作業は、一台の PC 上で行える。

今回は、作業時間の関係があり、当初から画像収集を検討せずにデータを取っているため、考察はできないのだが、画像を含めて網羅的にデータ収集できる余地もあるかもしれない。例えば、災害の記録を写真で伝えるブログなどの場合、画像も含めアーカイブされなければならない。

機関をまたいだブログアーカイブ計画を実行するとして、上記を総合し、必要な環境は、おそらく最小で次のようなものになるだろう。①必要な人手：プログラミングを担当する者（全体で数名～）、PC でプログラムを実行する者（各機関で少なくとも一人）。　②必要な機材：デスクトップ PC（各機関で少なくとも一台）、大容量記憶メディア（データ保存とバックアップに必要な数）。　③必要な時間：問題発生時適宜対応（プログラミング担当）、適宜のプログラム起動／終了と、不具合がないかの定期的な確認のためのそう長くない時間（プログラム実行者）。

6. まとめ

以上見てきたように、ブログデータのアーカイブは、社会的・研究的に大きな意義を有していると考えられるものの、その実践については議論すらはじまっていない状況にあると思われる。他方で、ブログサービスは営利目的で提供されており、アーカイブが行われなければ、将来的にはそのデータはサービス終了とともに消失してしまう。特に、近年ブログサービス終了の波が起こっており、アーカイブに関する議論は一刻も早く始められなければならない。

アーカイブに際しての課題は、純粋に技術的・法的なものではなく、議論を深めていく過程でクリアしていくべきものであり、アーカイブの必要性が理解されれば十分にクリア可能なものだと考えられる。またアーカイブを行うための環境を揃えるハードルは、アーカイブの重要性を鑑みれば、全く高くはない。研究者が音頭をとり、その重要性を周知するとともに、サービス提供者、アーカイブ実施機関と密に協力しながら、方法論を確立していく必要があるだろう。

■註

★01　ITmedia　20100331　「「CURURU」終了、ユーザーの移転先まとめページ公開」，
https://www.itmedia.co.jp/news/articles/1003/31/news086.html　（※最終閲覧 20221006）
★02　日経クロステック　20070420　「マイクロソフトが Windows Live スペースで SNS 機能を強化」，
https://xtech.nikkei.com/it/pc/article/NEWS/20070420/269091/　（※最終閲覧 20221006）
★03　ITmedia　20100928　「Microsoft、Windows Live のブログを閉鎖へ――WordPress に移行」，
https://www.itmedia.co.jp/enterprise/articles/1009/28/news031.html　（※最終閲覧 20221006）
★04　シーサー株式会社　20101224　「Seesaa ブログ、KDDI 株式会社の運営する au one ブログからの移行機能の提供を開始」，
https://www.seesaa.co.jp/news/article/179219763.html　（※最終閲覧 20221006）
★05　株式会社ベクター　20060530　「読者管理・課金機能付き拡張ブログサービス「Vector maglog」開始のお知らせ」
http://ir.vector.co.jp/corp/release/pr/20060530_1/　（※最終閲覧 20221006）
★06　あたしンちのおとうさんの独り言　20120218　「Vector maglog サービス終了」，
http://atasinti.chu.jp/dad3/archives/5523　（※最終閲覧 20221006）
★07　NTT コミュニケーションズ　20141201　「OCN ブログ人のサービス終了について」，
https://support.ntt.com/supportTopInfo/detail/pid25000004n3/　（※最終閲覧 20221006）
★08　ITmedia　20191216　「「Yahoo!ブログ」がサービス終了　「ネット上の遺産が消えて悲しい」「黒歴史が消えて安心した」などさまざまな声」，
https://www.itmedia.co.jp/news/articles/1912/16/news077.html　（※最終閲覧 20221006）
★09　まつゆう*/ 松丸祐子 メタバース DJ　20190803　「「ヤプログ！サービス終了」～プロデューサー時代の昔話を振り返る。」，
https://note.com/matsuyou/n/n896b8aeb94f1　（※最終閲覧 20221006）
★10　クルーズ株式会社 20080424　「「ＣＲＯＯＺ！ブログ」発！大人気ケータイ小説『同じ空の下で』が書籍化！！」，
https://prtimes.jp/main/html/rd/p/000000112.000000082.html　（※最終閲覧 20221006）
★11　クルーズ株式会社　20220602　「2005 年以来、30 億件以上のブログ記事を集め 500 万人以上のユーザーに利用された『CROOZ blog』2022 年 5 月をもってサービス終了へ」，
https://crooz.co.jp/post-12380　（※最終閲覧 20221006）
★12　ITmedia　20040701　「teacup、掲示板感覚のブログサイト「AutoPage」オープン」，
https://www.itmedia.co.jp/lifestyle/articles/0407/01/news101.html　（※最終閲覧 20221006）
★13　ITmedia　20220802　「「長い間ありがとう」　レンタル掲示板の草分け「teacup」終了 25 周年に」，

https://www.itmedia.co.jp/news/articles/2208/02/news082.html （※最終閲覧 20221006）
★14　ウェブリブログ事務局　20220118　「ウェブリブログサービス終了(2023/1)のお知らせ」，
https://info.at.webry.info/202201/article_2.html （※最終閲覧 20221006）
★15　気ままな旅日記　20190523　「Yahoo! ブログから WordPress への移行」
https://blog.zeke.jp/2019/05/23/9373/ （※最終閲覧 20221006）
★16　ホビーさんブログ 2　20190714　「コメントの引っ越しの件」，
https://yonyon199.blog.fc2.com/blog-entry-644.html （※最終閲覧 20221006）
★17　サイバーエージェント　掲載日記載なし　「沿革」，
https://www.cyberagent.co.jp/corporate/history/ （※最終閲覧 20221006）
★18　CROOZ blog　掲載日記載なし　「サービス終了のお知らせ」　，
https://blog.crooz.jp/ （※最終閲覧 20221006）
★19　Gloria 他［2012］ pp37.
★20　Luisa 他[2019] pp159.
★21　坂口[2012] pp292.
★22　同前 pp307
★23　Kyoung & Yi-Fan[2019]　※オンライン論文のためページ数なし
★24　「site:対象のドメイン名」で検索することでそのドメイン上の URL の概算が表示される
★25　プログラムを起動させ長期間の収集を行う場合、放熱性に優れたデスクトップ PC が推奨される
★26　日本図書館協会図書館の自由委員会　20110304　「岡崎市の図書館システムをめぐる事件について」，
https://www.jla.or.jp/portals/0/html/jiyu/okazaki201103.html （※最終閲覧 20221006）
★27　[site:]コマンドだけでは、トップページを含むすべてのページが表示される。URL の法則性を見つけ出し、[-][inurl:]などのコマンドを組み合わせることで、特定のページ群を検索できる場合がある
★28　国立国会図書館インターネット資料収集保存事業　20141001　「5. 収集する頻度」，
https://warp.da.ndl.go.jp/contents/reccommend/mechanism/mechanism05.html （※最終閲覧 20221006）
★29　参考までに、筆者の場合、プログラミング語言語初学者の状態から 20～30 時間程度（環境構築時間を含む）で、外部リストを読み込まない形でのスクレイピング（Python）を任意のブログに応用・外部ファイルにデータを取得することに成功した。もっとも、これは言語を理解したわけではなく、ウェブ上に多数存在する Python によるスクレイピングプログラムを「切り貼り」した上で調整し、目的の動作をさせる最低限のコツを掴んだに過ぎない。その後は、同じく生存学研究所スタッフで、中程度の知識を有する山口和紀のプロジェクト参加もあり、学習は中止している。

■文献

Gloria, MARK et al. 2012 “Blogs as a collective war diary”, *In: Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*: 37-46

Kyoung, LEE Tae; Yi-Fan, SU Leona 2019 “When a personal HPV story on a blog influences perceived social norms: the roles of personal experience, framing, perceived similarity, and social media metrics”, *Health communication*

Luisa, MARTINO Maria et al. 2019 “Cancer blog narratives: The experience of under-fifty women with breast cancer during different times after diagnosis”, *The Qualitative Report* 24-1: 158-173

坂口 由佳　2013　「自傷行為をする生徒たちに対して学校はどのような対応をしているのか――自傷行為経験者のブログから」，『教育心理学研究』61-3:290-310

Necessity of web log archiving, and tasks for its realization

Ryohei Nakai

Abstract:
This paper provides an overview of the significance of web log (hereafter: blog) archiving, and then examines the methods and issues involved. As a result, the following points became clear. (1) By examining how previous domestic and international studies that have analyzed narratives on blogs have considered the characteristics of them, it was found that the inherent value is placed on the narratives of people concerned on blogs in previous studies. (2) The authors' web-scraping efforts revealed that large-scale blog archives are technically feasible in an environment that can be prepared by individuals. (3) Based on the same efforts and discussions, it was concluded that the remaining legal and other issues can be sufficiently resolved through discussions on digital information archiving among the parties concerned.

Keyword:
Digital information archiving, weblogs (blogs), web scraping