CA1893

ウェブアーカイブの利活用に向けた動き ー世界の潮流と WARP の取組ー

前田直俊（まえだなおとし）*

1．はじめに

1990年代半ばにウェブアーカイブが行われ始めてから20年が経過し、その間、技術開発、法整備、運用構築、普及活動など様々な分野で取組が行われてきた。この数年はとりわけ利活用に向けた議論が活発になっている。本稿は、そうした動きと背景について概観するとともに、国立国会図書館インターネット資料収集保存事業（WARP）における利活用の取組を紹介する。

2．ウェブアーカイブの動向

2.1. IIPCに見る動向変化

ウェブアーカイブの動向を語る上で欠かせないのは国際インターネット保存コンソーシアム（International Internet Preservation Consortium：IIPC）の存在である。IIPCはウェブアーカイブに関する世界最大の国際的な組織体で、各国の国立図書館、アーカイブ機関、大学、研究機関など52機関が加盟している(1)。常設の運営委員会やワーキンググループからなる体制のもと、技術開発や共通課題の解決、ウェブサイトの共同収集などが行われているほか、年に１回開催されるIIPC総会では、加盟機関の関係者や研究者が一堂に会して課題の検討や情報交換が活発に行われる。そこで取り上げられる話題は、ウェブアーカイブの動向そのものであると言ってよい。

そこで、2011年から2015年までのIIPC総会における約200件のプレゼンテーション(2)を内容により９種類に分類して（表１）、その割合をグラフに示した（図１）。上位３種は「利活用」、「アーカイブ概要」、「技術」となっており、さらにこれら３種の割合について経年での推移を示したのが図２である。過去５年間において、「技術」が10％から20％と一定の割合を占めていること、また「利活用」の割合が15％から30％へと倍増した一方で、「アーカイブ概要」の割合が30％から０％へと大きく減少したことが見て取れる。

表１　IIPCプレゼンテーションの分類
（2011年から2015年の累積）

分類	内容	件数
利活用	ウェブアーカイブの利活用に関すること	43
アーカイブ概要	各機関におけるアーカイブの概要、収集戦略、収集状況、実践経験など	37
技術	クローラ、保存フォーマット、閲覧ソフト、長期保存などウェブアーカイブの技術に関すること	32
ワーキンググループ	IIPCに設置されているHarvesting、Access、Preservationなどのワーキンググループのセッション	30
IIPC運営事務	IIPCの運営や事務に関すること	25
プロジェクト	IIPCが正式に認定して実施するプロジェクトに関すること	10
コレクション構築	どのようなウェブサイトを収集するかなどコレクションの構築に関すること	7
法制度	ウェブアーカイブのための法整備に関すること	1
その他	その他	17
合計		202

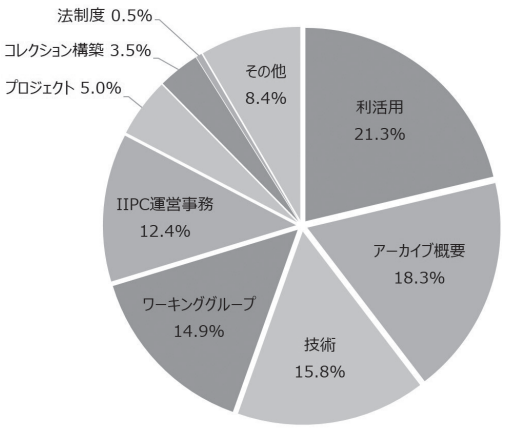

図１　IIPCプレゼンテーションの割合
（2011年から2015年の累積）

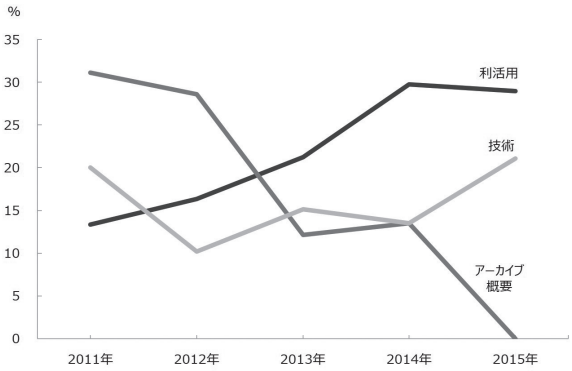

図２　IIPCプレゼンテーションの割合推移（上位３種）

2.2. 動向変化の背景

IIPCが設立された2003年は、世界各国でウェブアーカイブが本格的に実施され始めた時期であるが、技術面や運用面での整備がまだ十分になされていない状態であった。

技術については2000年代の半ばから後半にかけてIIPCの主導で開発が行われた結果、基礎技術の標準化と普及が進んだものの、進展するウェブ技術に対応するために恒常的な取組が必要となっている(3)。「技術」に関するプレゼンテーションが一定の割合を占めているのはそうした状況を反映していると言えよう。運用面については、各国における収集戦略や収集状況、実践経験などの「アーカイブ概要」について情報共有をしながら、課題の解決がなされてきた。

*関西館電子図書館課

こうして基礎技術の普及と運用の枠組み作りが進み、各国でアーカイブが形成されていく中で、相対的に重みを増してきたのが収集したコンテンツをいかに活用するかという視点である。プレゼンテーションにおける「利活用」と「アーカイブ概要」の割合変化はそうした動きの表れであり、2015年のIIPC総会においても「最近の議論の焦点はウェブアーカイブの利活用に移ってきている」という認識が示された（E1683参照）。このように、ウェブアーカイブはこれまでの「いかに集めるか」という段階から、「いかに活用するか」という段階へと移行していると言ってよい。

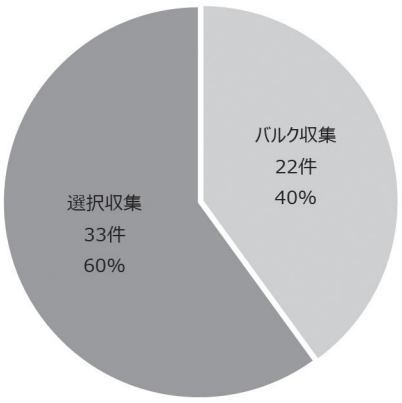

図３　バルク収集と選択収集の割合

3．閲覧利用とネット公開

3.1. 閲覧利用

ウェブアーカイブの典型的な使い方は過去のウェブサイトを閲覧することであり、消えてしまったウェブ情報を保存しているウェブアーカイブの価値が顕著に発揮される利用法である。

各アーカイブでは、同一のページを時間軸で遷移して閲覧できる機能を提供したり、URLやメタデータ、全文テキストによる検索機能を用意したりして[4]アクセスの利便を図っている。その他、リンク先のコンテンツが消えるなどしてリンク切れになった場合に、ウェブアーカイブに保存されているコンテンツにリンクすることでアクセスを保障するという使われ方もされている[5][6]。さらには、複数のウェブアーカイブを統合して閲覧できるようにするサービス[7]や、ウェブアーカイブを使って学術論文の引用文献のリンク切れを解決する取組[8]、利用者自らが必要に応じて特定コンテンツを保存して永続的にアクセス可能にするサービス[9][10]など、消えやすいウェブ情報へのアクセス手段としてウェブアーカイブは幅広く活用されている。

これらの利用法はアーカイブが自由にアクセスできる状態で公開されていることが前提となるが、ここで留意する必要があるのが、次節で述べるとおり、ウェブアーカイブで保存しているウェブサイトは必ずしもインターネットで公開できるものばかりではないという点である。

3.2. バルク収集と選択収集の公開範囲

ウェブアーカイブがインターネットで公開されるか否かの違いの多くは収集方法の違いに由来する。収集方法のうちの一つはバルク収集（Bulk harvesting）と呼ばれる方法で、国別コードトップレベルドメイン（例えば.ukや.fr）などのドメイン単位で包括的に集めるため、アーカイブの規模は極めて大きくなる。もう一つは選択収集（Selective harvesting）で、主題や出来事などに基づきウェブサイトを個別に選定して集めるため、比較的小規模なものが多い。

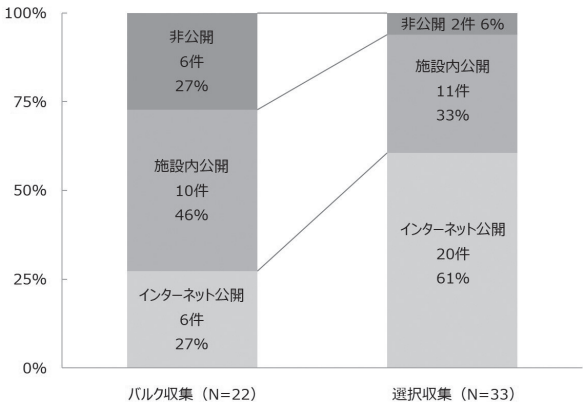

図４　公開範囲

図３はIIPC加盟機関のアーカイブのうち収集方法が判明している55件について、バルク収集と選択収集の割合を示したグラフである[11]。アーカイブの件数としては選択収集が上回っているものの、上述のようにバルク収集は規模が大きいため、保存しているコンテンツの量ではバルク収集の方が圧倒的に多くを占めると推測される。収集方法ごとの規模に関する全世界的な統計データは存在しないが、例えばバルク収集、選択収集の両方を実施している英国図書館（BL）のUK Web Archiveでは、2014年の１年間のバルク収集のデータ容量が56テラバイト[12]であったのに対し、選択収集のデータ容量は2004年から2016年までの13年間で28テラバイト[13]であり、年平均にすると２テラバイトとバルク収集の28分の１に過ぎない。

さらに、55件のアーカイブをバルク収集と選択収集に分けて、それぞれの公開範囲を示したのが図４である。バルク収集の多くは国立図書館が法制度に基づいて実施しており、サイト管理者の許諾を得ることなく包括的に収集できる反面、閲覧については著作権やプ

ライバシー保護などの観点から施設内公開や非公開と規定されている場合が多い。一方、選択収集では収集から公開までサイト管理者との間で権利処理を行うため、インターネット公開の割合が高くなっている。

4．研究利用に向けて

4.1. 研究利用への指向

このように、各国においてバルク収集により大規模な量のコンテンツが保存されているにも関わらず、それらの多くはインターネットで公開されない（できない）ため利用が極めて少ない[14]。長期的にはウェブ情報を文化遺産として後世に伝えるウェブアーカイブの意義は広く首肯されるものであるが、投資に見合う成果が短期的に現れにくい事業に対して理解が得られないケースもあり、中には予算の抑制や削減に直面するアーカイブもある[15]。

こうした背景のもと、実施機関の間では、ウェブアーカイブの利用価値についてより積極的にアピールして、潜在的な需要を掘り起こす必要性が強く認識されるようになった。そこで議論の焦点となってきたのが研究目的での利用促進である。ウェブアーカイブには膨大なウェブ情報が蓄積されており、ウェブ情報の分析で行われているデータ可視化やリンク解析、マッピングなどの手法を使うことで[16]、これまでにない新しい成果を生み出す可能性が秘められている。施設内公開や非公開のコンテンツについても、こうした二次的な使い方であれば著作権やプライバシー保護の観点からも問題はなく、利用対象に加えることができる。

4.2. 環境整備と研究者との連携

しかしながら、ウェブアーカイブを研究素材として使ってもらうのはそう簡単なことではない。BLが行った研究者からの聞き取り調査[17]によると、漠然と収集された膨大な量のデータの扱いにくさや、研究に適した分析ツールの欠如などから、ウェブアーカイブのコンテンツはそのままでは研究者にとって使いやすいものではないことが浮き彫りとなった。

そのため、アーカイブしたコンテンツからファイルフォーマット情報やリンク情報を抽出してデータセットとして公開したり[18]、コンテンツから分析用メタデータを抽出するための仕様を策定したりするなど[19]、ウェブアーカイブを研究素材として使えるようにするための環境整備が進められている。さらには研究者と共同で利活用の方法を探るプロジェクト[20]や研究者が使いやすい分析ツールの開発[21]、大規模なデータマイニングに適したプラットフォームの研究[22]なども行われている。

このように研究目的での利活用を進めていくためには、Dougherty[23]が指摘するとおり、研究者と積極的な連携を図りコミュニティを形成すること、分析ツールやデータセットなどの環境を整備すること、そして実践を積み重ねることが重要となってくる。

5．WARPにおける利活用の取組

5.1. WARPの現状

国立国会図書館（NDL）は2002年からインターネット資料収集保存事業（WARP）として日本国内のウェブサイトを対象としたウェブアーカイブを実施している[24]。2010年４月からは国立国会図書館法に基づき国の機関、地方自治体、独立行政法人、国公立大学などの公的機関のウェブサイトを網羅的に収集しているほか（E1046参照）、民間のウェブサイトについても、私立大学、政党、公益法人、学協会、第三セクター、業界団体、スポーツ団体、文化施設、国際的・文化的イベント、震災に関するものなど、公共性の高いサイトや社会的に有益なサイトを対象に選択収集を行っている。2016年12月現在の保存容量は860テラバイト、保存ファイル数は48億件であり、世界的にも有数の規模のアーカイブとなっている[25]。

5.2. WARPの公開範囲

WARPで収集したウェブサイトのうち公的機関のサイトについては、国立国会図書館法の規定によりNDLの館内で公開できるものの、公衆送信権に関する権利制限規定が設けられていないため、インターネットで公開するには権利者から許諾を得る必要がある[26]。そのため収集した全てのウェブサイトの管理者に対して個別に働きかけを行い、理解を得られたウェブサイトについてインターネット公開の許諾契約を取り交わしている。また民間のサイトについても、権利処理の際にインターネットでの公開について理解を求めており、その結果、WARPで保存しているコンテンツのうち85％がインターネット経由でアクセスできるようになっている。残りの15％はNDL館内で公開されており、非公開のものは存在しない。

このようにWARPは規模が大きくかつインターネット公開の割合が高いのが特徴であり、前章までに述べた利活用の観点に照らしてみると、閲覧利用から研究利用まで幅広い可能性を有しているアーカイブと言える。

5.3. WARPの利活用

WARPの閲覧利用については、ページビュー数にして月平均30万件のアクセスがあり、インターネットで公開しているコンテンツが多いことから、その90％以上がインターネット経由の利用となっている。

図５　WARPを使った分析、可視化の紹介サイト

中でもアクセス数が多いのは東京電力福島原子力発電所事故調査委員会（国会事故調）のウェブサイトである。2011年12月に福島第一原子力発電所事故の原因究明のために設置された同委員会は発足と同時にウェブサイトを公開し、事故調査に関する各種情報を発信していたが、2012年10月の事務局閉鎖とともにウェブサイトもインターネット上から消えてしまった。その後も原発事故に対する社会の関心は高く、事故調査をまとめた報告書がウェブサイトで公開されていたこともあって、WARPで保存している同サイトには依然として多数のアクセスがある[27]。消えたウェブ情報に対する需要の高さがうかがえる事例と言えよう。

また、単なる閲覧利用だけでなく、WARPをアーカイブ資源として積極的に活用する例も増えてきている。国の機関や地方自治体のウェブサイトにおいて古いコンテンツを削除する代わりに、WARPに保存されているコンテンツにリンクを張るという使い方である[28]。こうすることでサイト利用者が古い情報に永続的にアクセスできるだけでなく、サイト管理者にとってもサーバ容量や運用コストの削減につながるなど、双方にとって利点が大きい。公的機関のサイトを網羅的に保存しているWARPの特長を生かした使い方であり、多くの潜在的な需要が予想されるため、リンクの方法を紹介するページを設けたり[29]各機関に対して個別に案内をしたりして、積極的な働きかけを行っている。

さらにWARPのアーカイブデータを使った分析や可視化の試みも始まっている。WARPのウェブサイトでは、国の機関のウェブサイトの残存率調査（E1757参照）や都道府県ウェブサイト間のリンク関係の可視化、保存したウェブサイトの容量による可視化など、WARPを使って何ができるのかを端的にかつ分かりやすく伝えるコンテンツを公開している（図５）[30]。その他、WARPで保存している自治体サイトのデータを使って可視化を行うワークショップを開催するなど（E1840参照）、より多くの人に利用価値を知ってもらうことで多様な利活用の方法を探る試みも行っている。今後はデータ分析の専門家などの協力を得ながら、分析手法の確立やデータセットの公開に向けた検討を進めて行く予定である。

６．おわりに

以上見てきたように、各アーカイブが置かれた状況によって利活用の形態に違いはあるものの、その目指すところは、より多くの人にウェブアーカイブの価値を知ってもらい、使ってもらう、ということに尽きる。そのためには、環境の整備や方法論の確立に向けて引き続き多くの議論と実践を積み重ねていくとともに、ウェブアーカイブを使って何ができるのかを利用者に対して積極的に示していくことが求められる。

(1) 日本からは国立国会図書館が2008年に加盟し、これまでに全文検索エンジンNutchWAXの多言語対応（E995、E1185参照）やオリンピック・パラリンピック関連ウェブサイトの共同収集などで貢献を行っている。IIPCが共同収集したウェブサイトはオリンピック・パラリンピック関連も含め以下で公開されている。
International Internet Preservation Consortium. https://archive-it.org/home/IIPC, (accessed 2016-12-12).
(2) "General assembly". IIPC. http://netpreserve.org/general-assembly, (accessed 2016-12-12).
(3) 前田直俊ほか．ウェブアーカイブを支える技術．情報の科学と技術．2017，67(2)，p. 73-78.
(4) Costa, M. et al. The evolution of web archiving. International Journal on Digital Libraries. 2016. http://doi.org/10.1007/s00799-016-0171-9, (accessed 2016-12-12).
(5) AlNoamany, Yasmin. et al. Who and what links to the Internet Archive. International Journal on Digital Libraries. 2014, 14(3/4), p. 101-115. http://doi.org/10.1007/s00799-014-0111-5, (accessed 2016-12-12).
(6) Graham, M. "More than 1 million formerly broken links in English Wikipedia updated to archived versions from the Wayback Machine". Internet Archive. 2016-10-26. https://blog.archive.org/2016/10/26/more-than-1-million-formerly-broken-links-in-english-wikipedia-updated-to-archived-versions-from-the-wayback-machine/, (accessed 2016-12-12).
(7) Time Travel. http://timetravel.mementoweb.org/, (accessed 2016-12-12).
ロスアラモス国立研究所（米国）とオールド・ドミニオン大学（米国）が共同で実施しているMementoプロジェクト（前田．前掲）の一環として提供されている検索・閲覧サービス。
(8) Hiberlink. http://hiberlink.org/, (accessed 2016-12-12).
エディンバラ大学（英国）とロスアラモス国立研究所（米国）が共同で実施している。

(9) WebCite.
http://webcitation.org/, (accessed 2016-12-12).
雑誌の編集者や出版者などが参加するWebCite Consortiumが運営している。
(10) Perma.cc.
https://perma.cc/, (accessed 2016-12-12).
ハーバード大学（米国）が開発し、米国の約120の法律関係図書館がパートナーとして参加している（E1505参照）。
(11) 一つの機関でバルク収集と選択収集の両方を行っている場合は各1件として計上した。
(12) UK Web Archive. "2015 UK Domain Crawl has started". British Library. 2015-09-02.
http://blogs.bl.uk/webarchive/2015/09/2015-uk-domain-crawl-has-started.html, (accessed 2016-12-12).
(13) "UK Web Archive statistics". British Library.
http://www.webarchive.org.uk/ukwa/statistics, (accessed 2016-12-12).
(14) Hockx-Yu, H. "Web Archiving at National Libraries: Findings of Stakeholders' Consultation by the Internet Archive". Internet Archive. 2016-03.
https://archive.org/details/InternetArchiveStakeholdersConsultationFindingsPublic, (accessed 2016-12-12).
(15) Ibid.
(16) Meyer, E. T. et al. "Web archives: the future(s)". IIPC. 2011-06.
http://www.netpreserve.org/sites/default/files/resources/2011_06_IIPC_WebArchives-TheFutures.pdf, (accessed 2016-12-12).
(17) Hockx-Yu, H. "Up close and personal - Researchers and the UK Web Archive Project". IIPC. 2011-05.
http://gator1355.hostgator.com/~iipc/events/Hague/Presentations/Out%20of%20the%20Box/Researchers_HockxYu.pdf, (accessed 2016-12-12).
(18) "UK Web Archive Open Data". British Library.
http://data.webarchive.org.uk/opendata/, (accessed 2016-12-12).
(19) "Web Archive Transformation (WAT) Specification, Utilities, and Usage Overview". Internet Archive.
https://webarchive.jira.com/wiki/display/Iresearch/Web+Archive+Transformation+%28WAT%29+Specification%2C+Utilities%2C+and+Usage+Overview, (accessed 2016-12-12).
(20) "Big UK Domain Data for the Arts and Humanities". British Library.
http://buddah.projects.history.ac.uk/, (accessed 2016-12-12).
(21) "SHINE". British Library.
https://www.webarchive.org.uk/shine, (accessed 2016-12-12).
(22) Lin, Jimmy. "Warcbase: Building a Scalable Web Archiving Platform on Hadoop and HBase". IIPC. 2015-04.
http://netpreserve.org/sites/default/files/attachments/2015_IIPC-GA_Slides_15b_Lin.pptx, (accessed 2016-12-12).
(23) Dougherty, M. et al. "Researcher Engagement with Web Archives - State of the Art". JISC. 2010-08.
http://repository.jisc.ac.uk/544/, (accessed 2016-12-12).
(24) 国立国会図書館インターネット資料収集保存事業（WARP）.
http://warp.da.ndl.go.jp/, (参照 2016-12-12).
(25) 2014年時点で、世界の主要なウェブアーカイブのうちデータの保存容量が100テラバイトを超えるものは19％、保存ファイル数が10億件を超えるものは33％となっている（Coasta. Op. cit.）。
また、世界のウェブアーカイブ事業のリストによれば、WARPはデータ保存容量で第3位、保存ファイル数で第10位となっている。
"List of Web archiving initiatives".
https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives, (accessed 2016-12-12).
(26) 国立国会図書館総務部総務課ほか. インターネット資料の収集に向けて：国等の提供するインターネット資料を収集するための国立国会図書館法の改正について. 国立国会図書館月報. 2009, (581), p. 4-11.
http://doi.org/10.11501/1001142, (参照 2016-12-12).
(27) 国立国会図書館インターネット資料収集保存事業. "月間アクセスランキング". 国立国会図書館.
http://warp.da.ndl.go.jp/contents/ranking/index.html, (参照 2016-12-12).
(28) 例えば以下のページ。
"出版物等". 財務省.
http://www.mof.go.jp/jgbs/publication/, (参照 2016-12-12).
"埼玉県税務概況". 埼玉県.
http://www.pref.saitama.lg.jp/a0209/z-kurashiindex/z-gaikyou.html, (参照 2016-12-12).
(29) 国立国会図書館インターネット資料収集保存事業. "WARP活用術：古いページはWARPへリンク". 国立国会図書館.
http://warp.da.ndl.go.jp/contents/reccommend/utilization/warplink.html, (参照 2016-12-12).
(30) 国立国会図書館インターネット資料収集保存事業. "特色あるコレクション". 国立国会図書館.
http://warp.da.ndl.go.jp/contents/reccommend/collection/index.html, (参照 2016-12-12).

［受理：2017-02-06］

Naotoshi Maeda

The Movement toward Advancing the Use of Web Archives: Global Trends and the Use Cases of WARP