(26) 2005年ホワイトハウス・エイジング会議では、「図書館、生涯学習、情報と高齢者」というフォーラムが開かれた。“Pre-White House Conference on Aging Forum”. American Library Association. http://cs.ala.org/ra/whitehouse/, (accessed 2010-09-10).
(27) Turock, Betty J. “Libraries, Older Adults and the Future”. ALA Forum for the White House Conference on Aging. Chicago, Illinois, 2005-06-24, American Library Association. http://cs.ala.org/ra/whitehouse/WHCOAForumALA2005_turock.doc, (accessed 2010-09-10).
(28) 高島涼子. 高齢者生涯教育における図書館の役割. 京都大学生涯教育学 図書館情報学研究. 2005, (4), p. 195-202.
(29) Van Fleet, Connie. Libraries and Positive Aging: A Guide to Serving Older People. Libraries Unlimited, 2010, 200p.
(30) 前田章夫. 公共図書館における「2007年問題」. 図書館界. 2007, 59(2), p. 71-75.
(31) Dempsey, Beth. What boomers want. Library Journal. 2007, 132(12), p. 36-39.
(32) Rothstein, Pauline M. et al. Boomers and Beyond: Reconsidering the Role of Libraries. American Library Association, 2010, 152p.

CA1733

動向レビュー

ウェブアーカイブの課題と海外の取組み

1. はじめに

今日、ウェブはごく身近な情報源として一般的に利用されている。印刷された本や雑誌の形での発行が停止され、ウェブ上でのみ公開されるようになった刊行物も多い。今後、インターネットを利用しないと得られない情報は増加していくだろう。しかし、その速報性と更新のしやすさから、ウェブ上の情報は増加していくと同時に消失している。情報が載っていたページ自体が消失してしまうこともあるし、新しい情報が上書きされて古い情報が消失していることもある。ウェブ上のページにアクセスしようとして、「404 Not Found」というエラー画面を目にしたことも少なくないはずである。

ウェブページのURL（Uniform Resource Locators）の平均寿命は、44日から75日であると言われている(1)。このように失われやすい情報資源を、いかに後世に残すかが重要な課題となっている。

本稿ではこうした問題への取組みであるウェブアーカイブの概要と課題について説明したのち、海外の機関の取組みを紹介する。

2. ウェブアーカイブの概要と課題

ウェブアーカイブとは、インターネット上の情報を集め、将来の世代が利用できるように保存し、提供するサービスである。インターネット上から消失してしまった情報でも、ウェブアーカイブにより保存されていれば、永続的に見ることができるようになる。

多くのウェブアーカイブでは基本的にクローラーと呼ばれる自動収集プログラムによってデータを自動的・定期的に収集するか、作成者からデータを提供してもらうことで、情報を蓄積している。

ウェブアーカイブに保存される対象となるデータは、インターネットが発展すればするほど増加していく。最近は、ブログや動画共有サービスなど、だれもが情報の発信者になれるようなサービスが増加している。これらのコンテンツはひとつあたりの容量は小さいが、全体として膨大な量となっている。しかも更新が頻繁に行われ、情報が失われる速度も速いため、情報が更新される速度に収集する速度が追いつけなければ、情報に取りこぼしが生じる。このようなコンテンツを収集する場合、クローラーの技術的な問題や、作成者が多数存在していることによる著作権処理の問題など、解決すべき課題が多い。

収集した後にも課題は存在する。インターネット上の情報には、テキストや画像、動画、音声などがあり、それぞれのファイル形式は統一されておらず様々である。ある環境ではアクセスできる情報が、OSやブラウザの違い、バージョンの違いにより利用できないのでは問題である。また、たとえ膨大な量のウェブ情報を保存できたとしても、その中の有用な情報にたどり着く手段がなければ不十分である。利用者がどんな情報を必要としているのかを把握した上で、どのような提供の仕方をすれば利用者にとって使いやすいウェブアーカイブとなるのかを常に模索する必要がある。

3. 各国における取組み

ウェブアーカイブを構築している海外の機関[(2)]のうち、特色ある取組みを行っている機関をいくつか紹介する。

3-1. Internet Archive

Internet Archive（以下IA）は1996年に設立された米国の非営利法人で、2007年にはカリフォルニア州から図書館として認定された[(3)]。IAは全世界を対象としてウェブ情報の収集を行っている。すなわち、管理者がアクセス制限を行っている場合を除き、インターネットで公開されている情報すべてが収集の対象となっている。IAのデータベースの規模はインターネットの拡大とともに成長し、毎月圧縮ファイルで100テラバイト近くの容量が増加している[(4)]。これらの情報は“Wayback Machine”という閲覧システムによってインターネットで公開されており、自由に閲覧が可能である。

また、IAでは“Archive-It”という有料サービスを提供している[(5)]。これは、専門的な知識がない機関でも、自組織のウェブサイトのアーカイブが行えるようにするサービスで、サービスの契約者によって指定されたデータをIAが収集し、メタデータを作成し、全文検索ができるようにするものである。米国の政府機関や大学など、様々な機関がこのサービスを利用して自組織のウェブサイトのアーカイブを実施している[(6)]。特定の組織だけでなく、様々な機関においてウェブアーカイブへの意識が高まれば、より有益な情報が将来の世代のために保存できるだろう。

3-2. 米国議会図書館

米国議会図書館（Library of Congress；以下LC）で行われているウェブアーカイブでは、前項のIAにおいて収集されたウェブコンテンツを、選択的に整備して公開している[(7)]。例えば、米国大統領選挙、9.11同時多発テロ、イラク戦争といったテーマごとに分類されている。なお、管理者の許諾に基づき提供しているため、管理者の要望があればコンテンツの公開が制限されることもある。

他に、LCでは、“K-12 Web Archiving”というプロジェクトを行っている[(8)]。このプロジェクトはLCがIA、カリフォルニア電子図書館（California Digital Library）と協同で実施しているもので、プロジェクト参加校の小学生、中学生、高校生が、将来のために残したいと思うウェブサイトを、テーマを決めていくつか選び、前述の“Archive-It”を利用して保存するものである。子どもの目線をウェブサイトの選定基準に取り入れるとともに、ウェブ情報を保存する取組みを一般に周知する上で、有効であると考えられる。

さらに、2010年にはソーシャルメディアの1つであるツイッター（Twitter）から、公開設定となっているツイート（ツイッターに投稿された140字以内のメッセージ）すべてがLCに寄贈された（E1042参照）。2006年にツイッターのサービスが開始してからこれまでに投稿されたツイートはもちろん、今後増えていくツイートも収集の対象となる。収集されたツイートは一般公開の予定はないが、研究目的での利用が可能である。この寄贈によって、他の機関においても、データの寄贈という手段が、ウェブアーカイブにおける一つの手法であるということが印象付けられたのではないだろうか。

3-3. 英国国立公文書館

英国国立公文書館（The National Archives；以下TNA）では、“UK Government Web Archive”というウェブアーカイブのプロジェクトを行っている[(9)]。これは英国の政府機関のウェブサイトを保存しているもので、現在は年に3回の頻度で収集を行っている。また、このプロジェクトで保存されたデータを利用して、“Web Continuity”というサービスを行っている。このサービスでは、まず、政府機関のウェブサイトの管理者に依頼して、各ウェブサイトに簡単なソフトウェアをインストールしてもらう。そうすることで、もし利用者が政府機関のウェブサイトから何か情報を得ようとした際に、そのページが現在利用できない状態となっていたとしても、TNAのウェブアーカイブで保存されているページであれば、自動的にそのページに転送されるように設定できる。すなわち、ページが移動していたり、削除されていたりしたとしても、「404 Not Found」というエラー画面を表示させないで、任意のページを表示させる

ことができるようになる。また、転送が行われた先のページには、TNAにより保存されたページであることがわかるような注意書きが挿入されているため、現在の情報と過去の情報は明確に区別できるように工夫がされている。2008年11月にサービスが開始されてから、2009年1月に80万件だったTNAのウェブアーカイブへのアクセス数は、同年10月には900万件に増加したという(10)。この取組みは、ウェブアーカイブにおける可視性の向上に有用であると考えられる。

3-4. 英国図書館

英国図書館（British Library；以下BL）では、英国ドメイン（.uk）のウェブサイトを管理者の許諾を得て選択的に収集している(11)。収集や保存は“Web Curator Tool”(12)（以下WCT）というシステムを利用して管理されており、収集されたコンテンツは2010年2月から“UK Web Archive”で正式に公開されている。

BLは、2,400時間に及ぶ映像を公開していた、民間の大規模なウェブサイト“One & Other”の保存について報告している(13)。このウェブサイトでは2009年7月から10月にかけて、2,400人の一般の人々の日常生活（1人1時間、計2,400時間分）の映像が、ストリーミングメディアの形式で公開されていた。ストリーミングメディアのようなリアルタイムで再生する形式のコンテンツをクローラーで収集するのは困難である。実際、クローラーとしてHeritrix（CA1664参照）を使用しているWCTで収集を試みたところ、ウェブページを構成するHTMLファイルの収集しかできなかった。当該ウェブサイトは公開終了がせまっていたため、BLはストリーミングメディアをファイルの形で保存できるJaksta(14)というソフトウェアを使用して、映像を保存した。元々のファイルに異常があったものは正常には保存ができなかったが、スポンサーと費用を分担してファイルの修復がなされる予定である。現在公開が終了している“One & Other”のウェブサイトを閲覧しようとすると、直接“UK Web Archive”内のサイトに転送されるようになっている。“One & Other”は公開当時に注目されていたウェブサイトであったため、それを保存したことで、“UK Web Archive”への関心も集めることとなった。収集・保存が困難なウェブサイトも、コストの分担やデータの提供など、関係者の協力があれば保存が可能である。

3-5. IIPC

最後に、ウェブアーカイブの国際連携を目的とした組織である、国際インターネット保存コンソーシアム（International Internet Preservation Consortium；以下IIPC；CA1664参照）の活動を紹介する。IIPCは2003年に結成され、2010年9月時点で各国の国立図書館など39機関から構成されている(15)。日本の国立国会図書館（以下NDL）は2008年4月にIIPCに加盟し、IIPCが開発したウェブアーカイブ用ツールの多言語対応等に取り組んでいる。

IIPCでは、クローラーの開発や、保存形式など規格の標準化、検索システムの検討など、ウェブアーカイブにおける様々な課題について協調して取り組んでいる。そして、定期的に総会を開催し、各機関のウェブアーカイブの事例や開発成果を共有している。

例えば、2010年5月にシンガポールで開催された総会では、ロスアラモス国立研究所（Los Alamos National Laboratory）により、現在のウェブサイトと過去のウェブサイトを一体化するプロジェクト“Memento”について報告された(16)。このプロジェクトで研究されているプラグインを自分のブラウザに導入することで、その都度ウェブアーカイブのサイトに移動しなくても過去のページを閲覧できるようになる、というものである。現在のウェブサイトでリンクが切れているページが選択されたとき、“Time Gate”と呼ばれる機能により、過去のページが表示されるようになる。日付を選択できるスクロールバーがあるので、ニュースサイトのような更新が頻繁に行われるウェブサイトで、どのようにページが更新されていったのかを簡単に見ることができる。この機能が一般化されれば、過去のウェブサイトをより手軽に閲覧できるようになり、ウェブ情報の起源や、情報が遷移する過程を容易に理解できるようになるだろう。Mementoプロジェクトは発表後、IIPCの正式タスクとするよう提案があり、BLやLC、オランダ王立図書館などの支援を受けることとなった(17)。

このようにIIPCでは、各機関が他機関と連携することで、知識や情報を共有し、課題解決に取り組んでいる。

インターネットというのは、国や地域の境界を越えた広がりを持っている。そのような情報を収集・整備し、課題を解決していくためには、国際的な連携が必要だという結論に行きつくのは自然なことである。

4. おわりに

日本では、2010年4月から、NDLによる、公的機関のウェブサイトの包括的収集が開始された（E1046参照）。収集のためには、対象機関の理解と協力が必

要であり、さらなる広報の充実が課題となっている。

また、収集範囲が増大したことで、収集した大量のデータを保存する領域の確保も大きな問題である。このため、データ量の削減を図るべく、前回収集時から変更のあったファイルだけを収集・保存し、提供する差分収集の仕組みの導入を検討している。差分収集したデータの提供に関しては、ウェブアーカイブ閲覧用ソフトウェア“Wayback”の適用可能性について2010年度に調査を行っている。Waybackは3-1で紹介した“Wayback Machine”を、IIPCがOSS（オープンソースソフトウェア）版として改良したシステムである。差分収集データ提供機能として、変更がなかったため収集されなかったウェブページに閲覧要求がなされた際、要求された日時から最も近い日時に収集されたデータを表示することができる。このようなIIPCの成果物の、NDLのウェブアーカイブへの導入にあたっては、必要に応じてIIPCと協議しながら行っていくことになるだろう。

ウェブアーカイブを発展させるためには、技術的な面を含め、国際的な連携が不可欠である。2011年11月には、IIPCのワーキンググループ会議の開催がつくば市で予定されている（E1109参照）。さらなる連携の進展を期待したい。

（関西館電子図書館課：中島美奈（なかじまみな））

(1) Guy, Marieke. “What's the average lifespan of a Web page?”. JISC-PoWR. 2009-08-12. http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page/, (accessed 2010-11-11).
(2) 代表的なもののリストが下記のウェブページにある。“Member Archives”. international internet preservation consortium. http://netpreserve.org/about/archiveList.php, (accessed 2010-11-11).
(3) “Internet Archive Forums: View Post”. Internet Archive. 2007-06-25. http://www.archive.org/post/121377/internet-archive-officially-a-library, (accessed 2010-11-11).
(4) “Member Archives”. international internet preservation consortium. http://netpreserve.org/about/archiveList.php, (accessed 2010-11-11).
(5) Archive-It. http://www.archive-it.org/, (accessed 2010-11-11).
(6) “Partners”. Archive-It. http://www.archive-it.org/public/partners.html, (accessed 2010-11-11).
(7) “Web Archiving”. Library of Congress. http://www.loc.gov/webarchiving/index.html, (accessed 2010-11-11).
(8) “K-12 Web Archiving”. Archive-It. http://www.archive-it.org/k12/, (accessed 2010-11-11).
(9) “COI - Appendix A UK Government Web Archive”. The Central Office of Information. http://coi.gov.uk/guidance.php?page=245, (accessed 2010-11-11).
(10) “Web continuity”. The National Archives. http://www.nationalarchives.gov.uk/information-management/policies/web-continuity.htm, (accessed 2010-11-11).
(11) “UK Web Archive: About”. UK Web Archive. http://www.webarchive.org.uk/ukwa/info/about, (accessed 2010-11-11).
(12) Web Curator Tool. http://webcurator.sourceforge.net/, (accessed 2010-11-11).
(13) Hockx-Yu, Helen et al. “Capturing and Replaying Streaming Media in a Web Archive – A British Library Case Study”. iPRES 2010. http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/hockxyu-44.pdf, (accessed 2010-11-11).
(14) Jaksta. http://www.jaksta.com/, (accessed 2010-11-11).
(15) “Members ”. international internet preservation consortium. http://netpreserve.org/about/memberList.php, (accessed 2010-11-11).
(16) Sanderson, Robert et al. “Memento: Integrating the Past and Current Web”. international internet preservation consortium. http://netpreserve.org/events/2010GApresentations/memento_pres_opt.pdf, (accessed 2010-11-11).
(17) Hockx-Yu, Helen et al. “Access Working Group – Final Report of Activities”. international internet preservation consortium. http://netpreserve.org/events/2010GApresentations/03a_IIPC_AWG_FinalReport_SingaporeGA.pdf, (accessed 2010-11-11).

Ref:
柊和佑ほか. 特集, Webアーカイビングの現状と課題: 世界のWebアーカイブ－IIPC（International Internet Preservation Consortium）を中心にして. 情報の科学と技術. 2008, 58(8), p. 389-393.
武田和也. 特集, Webアーカイビングの現状と課題: 海外動向との対比からみた日本のWebアーカイビングの課題と展望―国立国会図書館の取り組みを通して―. 情報の科学と技術. 2008, 58(8), p. 394-400.